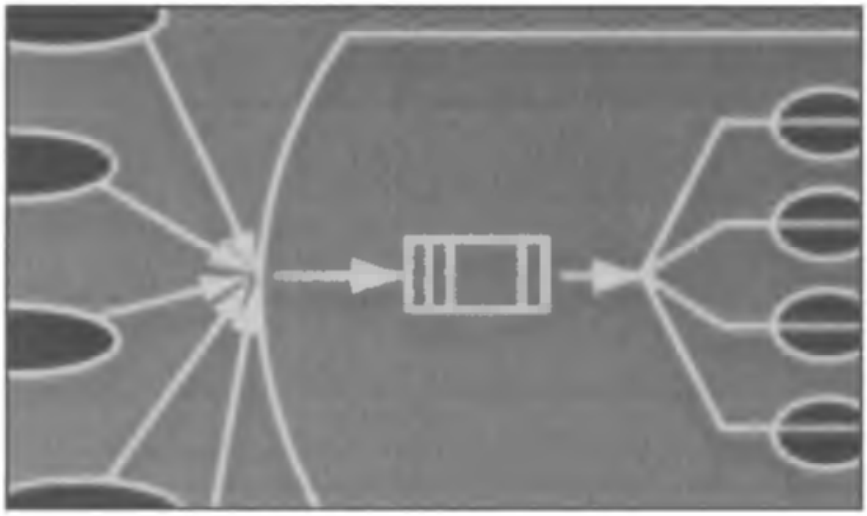
STOCHASTIC MODELS IN QUEUEING THEORY

Second Edition

MEDHI

STOCHASTIC MODELS IN QUEUEING THEORY

J. MEDHI



STOCHASTIC MODELS IN QUEUEING THEORY

SECOND EDITION

ACADEMIC PRESS
An imprint of Elsevier Science

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Senior Sponsoring Editor	Barbara Holland
Project Manager	Nancy Zachor
Editorial Coordinator	Tom Singer
Cover Design	Shawn Girsberger
Copyeditor	Deborah Prato
Composition	International Typesetting and Composition
Printer	Maple-Vail

This book is printed on acid-free paper. ♾

Copyright 2003, 1991, Elsevier Science (USA)

All rights reserved.
No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Requests for permission to make copies of any part of the work should be mailed to: Permissions Department, Harcourt, Inc., 6277 Sea Harbor Drive, Orlando, Florida 32887-6777.

Academic Press
An imprint of Elsevier Science
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

Academic Press
An imprint of Elsevier Science
200 Wheeler Road, Burlington, Massachusetts 01803, USA
http://www.academicpress.com

Academic Press
An imprint of Elsevier Science
84 Theobald's Road, London WC1X 8RR, UK
http://www.academicpress.com

Library of Congress Control Number: 2002110814
International Standard Book Number: 0-12-487462-2

PRINTED IN THE UNITED STATES OF AMERICA
03 04 05 06 9 8 7 6 5 4 3 2

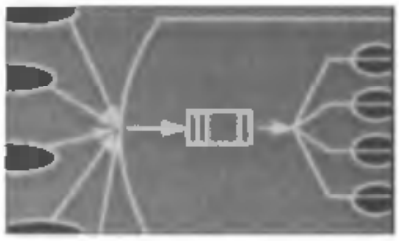
In the sacred memory
of my parents,
Binandi Chandra
&
Kadambari

karmaṇya evā 'dhikāras te
    mā phaleṣu kadācana
mā karmaphalahetur bhūr
    mā te saṅgo'stv akarmaṇi

To action alone thou hast a right and never at all to the fruits; let not fruits of action be thy motive; neither let there be in thee any attachment to inaction.

The Bhagavadgītā, II, 47

Contents



Preface xv

CHAPTER 1 Stochastic Processes 1

1.1 Introduction 1

1.2 Markov Chains 2

1.2.1 Basic ideas 2

1.2.2 Classification of states and chains 4

1.3 Continuous-Time Markov Chains 14

1.3.1 Sojourn time 14

1.3.2 Transition density matrix or infinitesimal generator 15

1.3.3 Limiting behavior: ergodicity 16

1.3.4 Transient solution 18

1.3.5 Alternative definition 19

1.4 Birth-and-Death Processes 23

1.4.1 Special case: $M/M/1$ queue 25

1.4.1 Pure birth process: Yule-Furry process 25

1.5 Poisson Process 25

1.5.1 Properties of the Poisson process 28

1.5.2 Generalization of the Poisson process 29

1.5.3 Role of the Poisson process in probability models 31

1.6 Randomization: Derived Markov Chains 32

1.6.1 Markov chain on an underlying Poisson process (or subordinated to a Poisson process) 33

1.6.2 Equivalence of the two limiting forms 33
1.6.3 Numerical method 34
1.7 Renewal Processes 35
1.7.1 Introduction 35
1.7.2 Residual and excess lifetimes 36
1.8 Regenerative Processes 37
1.8.1 Application in queueing theory 38
1.9 Markov Renewal Processes and Semi-Markov Processes 39
Problems 41
References and Further Reading 46

CHAPTER 2 Queueing Systems: General Concepts 47

2.1 Introduction 47
2.1.1 Basic characteristics 48
2.1.2 The input or arrival pattern of customers 48
2.1.3 The pattern of service 49
2.1.4 The number of servers 49
2.1.5 The capacity of the system 49
2.1.6 The queue discipline 49
2.2 Queueing Processes 50
2.3 Notation 51
2.4 Transient and Steady-State Behavior 52
2.5 Limitations of the Steady-State Distribution 53
2.6 Some General Relationships in Queueing Theory 54
2.7 Poisson Arrival Process and Its Characteristics 59
2.7.1 PASTA: Poisson arrivals see time averages 59
2.7.2 ASTA: arrivals see time averages 62
References and Further Reading 62

CHAPTER 3 Birth-and-Death Queueing Systems: Exponential Models 65

3.1 Introduction 65
3.2 The Simple $M/M/1$ Queue 65
3.2.1 Steady-state solution of $M/M/1$ 66
3.2.2 Waiting-time distributions 68
3.2.3 The output process 72
3.2.4 Semi-Markov process analysis 75
3.3 System with Limited Waiting Space: The $M/M/1/K$ Model 77
3.3.1 Steady-state solution 77
3.3.2 Expected number in the system L_K 78
3.3.3 Equivalence of an $M/M/1/K$ model with a two-stage cyclic model 80

3.4 Birth-and-Death Processes: Exponential Models 81
3.5 The $M/M/\infty$ Model: Exponential Model with an Infinite Number of Servers 83
3.6 The Model $M/M/c$ 84
3.6.1 Steady-state distribution 84
3.6.2 Expected number of busy and idle servers 87
3.6.3 Waiting-time distributions 89
3.6.4 The output process 93
3.7 The $M/M/c/c$ System: Erlang Loss Model 95
3.7.1 Erlang loss (blocking) formula: Recursive algorithm 99
3.7.2 Relation between Erlang's B and C formulas 100
3.8 Model with Finite Input Source 101
3.8.1 Steady-state distribution: $M/M/c//m$ $(m > c)$. Engset delay model 101
3.8.2 Engset loss model $M/M/c//m/(m > c)$ 106
3.8.3 The model $M/M/c//m$ $(m \leq c)$ 109
3.9 Transient Behavior 110
3.9.1 Introduction 110
3.9.2 Difference-equation technique 112
3.9.3 Method of generating function 117
3.9.4 Busy-period analysis 119
3.9.5 Waiting-time process: Virtual waiting time 125
3.10 Transient-State Distribution of the $M/M/c$ Model 127
3.10.1 Solution of the differential-difference equations 127
3.10.2 Busy period of an $M/M/c$ queue 133
3.10.3 Transient-state distribution of the output of an $M/M/c$ queue 136
3.11 Multichannel Queue with Ordered Entry 138
3.11.1 Two-channel model with ordered entry (with finite capacity) 139
3.11.2 The case $M = 1, N = N$ 140
3.11.3 Particular case: $M = N = 1$ (overflow system) 142
3.11.4 Output process 144
Problems and Complements 145
References and Further Reading 159

CHAPTER 4 Non-Birth-and-Death Queueing Systems: Markovian Models 165

4.1 Introduction 165
4.1.1 The system $M/E_k/1$ 165
4.1.2 The system $E_k/M/1$ 170

4.2 Bulk Queues 174
4.2.1 Markovian bulk-arrival system: $M^X/M/1$ 174
4.2.2 Equivalence of $M^r/M/1$ and $M/E_r/1$ systems 178
4.2.3 Waiting-time distribution in an $M^X/M/1$ queue 178
4.2.4 Transient-state behavior 179
4.2.5 The system $M^X/M/\infty$ 181
4.3 Queueing Models with Bulk (Batch) Service 185
4.3.1 The system $M/M(a, b)/1$ 186
4.3.2 Distribution of the waiting-time for the system $M/M(a, b)/1$ 190
4.3.3 Service batch-size distribution 195
4.4 $M/M(a, b)/1$: Transient-State Distribution 196
4.4.1 Steady-state solution 198
4.4.2 Busy-period distribution 198
4.5 Two-Server Model: $M/M(a, b)/2$ 202
4.5.1 Particular case: $M/M(1, b)/2$ 204
4.6 The $M/M(1, b)/c$ Model 205
4.6.1 Steady-state results $M/M(1, b)/c$ 208
Problems and Complements 210
References and Further Reading 217

CHAPTER 5 Network of Queues 221

5.1 Network of Markovian Queues 221
5.2 Channels in Series or Tandem Queues 222
5.2.1 Queues in series with multiple channels at each phase 224
5.3 Jackson Network 226
5.4 Closed Markovian Network (Gordon and Newell Network) 233
5.5 Cyclic Queue 236
5.6 BCMP Networks 238
5.7 Concluding Remarks 240
5.7.1 Loss networks 241
Problems and Complements 242
References and Further Reading 249

CHAPTER 6 Non-Markovian Queueing Systems 255

6.1 Introduction 255
6.2 Embedded-Markov-Chain Technique for the System with Poisson Input 256

6.3 The $M/G/1$ Model: Pollaczek-Khinchin Formula 259
6.3.1 Steady-state distribution of departure epoch system size 259
6.3.2 Waiting-time distribution 261
6.3.3 General time system size distribution of an $M/G/1$ queue: supplementary variable technique 267
6.3.4 Semi-Markov process approach 274
6.3.5 Approach via martingale 274
6.4 Busy Period 276
6.4.1 Introduction 276
6.4.2 Busy-period distribution: Takács integral equation 277
6.4.3 Further discussion of the busy period 279
6.4.4 Delay busy period 284
6.4.5 Delay busy period under N-policy 285
6.5 Queues with Finite Input Source: $M/G/1//N$ System 289
6.6 System with Limited Waiting Space: $M/G/1/K$ System 292
6.7 The $M^X/G/1$ Model with Bulk Arrival 295
6.7.1 The number in the system at departure epochs in steady state (Pollaczek-Khinchin formula) 295
6.7.2 Waiting-time distribution 295
6.7.3 Feedback queues 302
6.8 The $M/G(a, b)/1$ Model with General Bulk Service 304
6.9 The $G/M/1$ Model 306
6.9.1 Steady-state arrival epoch system size 306
6.9.2 General time system size in steady state 309
6.9.3 Waiting-time distribution 311
6.9.4 Expected duration of busy period and idle period 313
6.10 Multiserver Model 314
6.10.1 The $M/G/\infty$ model: transient-state distribution 314
6.10.2 The model $G/M/c$ 319
6.10.3 The model $M/G/c$ 322
6.11 Queues with Markovian Arrival Process 324
Problems and Complements 326
References and Further Reading 334

CHAPTER 7 Queues with General Arrival Time and Service-Time Distributions 339

7.1 The $G/G/1$ Queue with General Arrival Time and Service-Time Distributions 339
7.1.1 Lindley's integral equation 341
7.1.2 Laplace transform of W 343
7.1.3 Generalization of the Pollaczek-Khinchin transform formula 346
7.2 Mean and Variance of Waiting Time W 348
7.2.1 Mean of W (single-server queue) 348
7.2.2 Variance of W 351
7.2.3 Multiserver queues: approximation of mean waiting time 353
7.3 Queues with Batch Arrivals $G^{(x)}/G/1$ 356
7.4 The Output Process of a $G/G/1$ System 358
7.4.1 Particular case 359
7.4.2 Output process of a $G/G/c$ system 360
7.5 Some Bounds for the $G/G/1$ System 360
7.5.1 Bound for $E(I)$ 360
7.5.2 Bounds for $E(W)$ 360
Problems and Complements 368
References and Further Reading 371

CHAPTER 8 Miscellaneous Topics 375

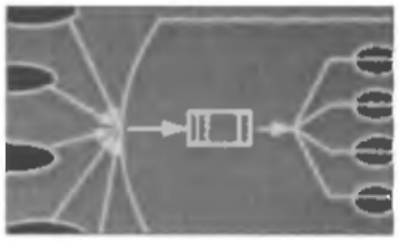
8.1 Heavy-Traffic Approximation for Waiting-Time Distribution 375
8.1.1 Kingman's heavy-traffic approximation for a $G/G/1$ queue 375
8.1.2 Empirical extension of the $M/G/1$ heavy-traffic approximation 379
8.1.3 $G/M/c$ queue in heavy traffic 381
8.2 Brownian Motion Process 383
8.2.1 Introduction 383
8.2.2 Asymptotic queue-length distribution 386
8.2.3 Diffusion approximation for a $G/G/1$ queue 389
8.2.4 Virtual delay for the $G/G/1$ system 391
8.2.5 Approach through an absorbing barrier with instantaneous return 394
8.2.6 Diffusion approximation for a $G/G/c$ queue: state-dependent diffusion equation 395

8.2.7 Diffusion approximation for an $M/G/c$ model 396
8.2.8 Concluding remarks 397
8.3 Queueing Systems with Vacations 398
8.3.1 Introduction 398
8.3.2 Stochastic decomposition 399
8.3.3 Poisson input queue with vacations: [exhaustive-service] queue-length distribution 399
8.3.4 Poisson input queue with vacations: waiting-time distribution 404
8.3.5 $M/G/1$ system with vacations: nonexhaustive service 406
8.3.6 Limited service system: $M/G/1-V_m$ model 407
8.3.7 Gated service system: $M/G/1-V_m$ model 408
8.3.8 $M/G/1/K$ queue with multiple vacations 412
8.3.9 Mean value analysis through heuristic treatment 416
8.4 Design and Control of Queues 423
8.5 Retrial Queueing System 427
8.5.1 Retrial queues: model description 427
8.5.2 Single-server model: $M/M/1$ retrial queue 429
8.5.3 $M/G/1$ retrial queue 432
8.5.4 Multiserver models 436
8.5.5 Model with finite orbit size 439
8.5.6 Other retrial queue models 440
8.6 Emergence of a New Trend in Teletraffic Theory 441
8.6.1 Introduction 441
8.6.2 Heavy-tail distributions 442
8.6.3 $M/G/1$ with heavy-tailed service time 445
8.6.4 Pareto mixture of exponential (PME) distribution 445
8.6.5 Gamma mixture of Pareto (GMP) distribution 447
8.6.6 Beta mixture of exponential (BME) distribution 450
8.6.7 A class of heavy-tail distributions 452
8.6.8 Long-range dependence 454
Problems and Complements 455
References and Further Reading 461

Appendix 469
Index 477

This Page Intentionally Left Blank

Preface



Overview

The study of queueing models has been of considerable active interest ever since the birth of queueing theory at the beginning of the last century. Queueing theory continues to be one of the most extensive theories of stochastic models. Its progress and development, both in methodology and in applications, are ever growing. Innovative analytic treatments toward its theoretical development are being advanced, and newer areas of application are emerging.

There is a large and growing audience interested in the study of queueing models. The level of background and preparation among them varies a great deal, along with their requirements for depth of coverage. The audience is composed of advanced undergraduate and graduate students from a number of disciplines. In addition to students of standard graduate courses, there are many researchers, professionals, and industry analysts who require an in-depth knowledge of the subject.

There are, of course, some excellent advanced works, monographs, and texts on the subject as well. The rapid development of the subject demands updated texts, especially for the type of audience aimed at. Furthermore, the style of presentation and the approach of individual authors appeal to different sections of this large and varied audience.

The author feels that there is sufficient scope and material to warrant additional texts, especially at the graduate level, in this ever-growing subject area. This book has grown out of the author's long experience of teaching and research in India, the United States, and Canada. A reviewer's glowing compliment (in *American Mathematical Monthly*) on the author's first book *Stochastic Processes* (Wiley Eastern, and Halsted Wiley 1982) inspired the author to undertake preparation of a book on queueing models in a similar readable style.

Organization of the book

The book is divided into eight chapters. Chapter 1 is a summary of basic results in stochastic processes. This should be helpful to users in eliminating the need to refer frequently to other books on stochastic processes just for basic results. Chapter 2, which is devoted to general concepts, contains some discussions on concepts such as PASTA, superposition of arrival processes, and customer and time averages. Chapters 3 and 4 deal with birth-and-death queueing models and non-birth-and-death systems, respectively. Transient behavior and busy period analysis have been discussed at some length, and a uniformity of approach is emphasized. Some models of bulk queues have also been included because of their importance in transportation science. Chapter 5 is devoted to networks of queues and Chapter 6 to certain non-Markovian queueing systems. In Chapter 7, systems with both general arrival and service patterns are discussed. Chapter 8 covers miscellaneous topics such as asymptotic methods and queues with vacations, with a brief excursion into the design and control of queues. Diffusion approximations, which have emerged as powerful tools, have been discussed in some detail. We believe this chapter will be especially useful to researchers and professionals who wish to have a broad, general idea of the diffusion approximation methods.

Each of the chapters (except Chapter 2) contains a number of worked examples and problems, and all the chapters include extensive and recent references. The problems contain some materials that have been discussed, keeping in mind researchers and those who wish to pursue the subject further.

Changes to the new edition

In order to facilitate use of the second edition by those who are already familiar with the first edition, a drastic change in the basic structure has been avoided. The number of chapters has been kept at eight, with considerable additions in the broad topics mainly based on recent developments during the intervening years. Apart from inclusion of new topics (including some emerging during the past few years), new examples, and new problems, topical discussions have been expanded through notes, remarks, and so on. References have been updated. These have been supplemented by related works of interest for further reading. Chapters 3, 6, and 8, in particular, contain many new topics. Some of the new matters address finite input source and finite buffer models, advanced vacation models, retrial queueing systems, and a newly emerging trend in teletraffic processes and their analyses. My sincere hope is that the book will be found useful as a graduate text and also as a reference book by professionals and researchers in this subject area.

In addition to mathematics and statistics, the book could be used for a one- or two-semester course at the advanced undergraduate or graduate level in operations research, computer science, systems science, industrial (and other

branches of) engineering, telecommunications, economics, management, and business (with programs focused on quantitative methods).

Course coverage

The prerequisites for using this book are a course on applied probability and a course on advanced calculus.

Teachers would be the best judges of topics to be covered in a course. The following suggestions are for their consideration:

For a two-semester course:
The whole book.
For a one-semester course:
Sections 1.1 through 1.5; 1.7 through 1.9;
Sections 2.1 through 2.7;
Sections 3.1 through 3.8 and 3.11;
Sections 4.2 and 4.3;
Sections 5.1 through 5.4;
Sections 6.1 through 6.4; 6.7, 6.9 and 6.10; and
Sections 7.1 and 8.1
Exercises are to be selected from problems and complements.

Acknowledgments

I am intellectually indebted to all those whose works have stimulated my interest in this subject area. I have drawn freely and widely from the ever-increasing body of literature.

In preparing this book, I have received encouragement and assistance in various ways from a number of experts, friends, and colleagues from this country and abroad. I am thankful to them all.

I am most grateful to Professor J. G. C. Templeton (University of Toronto) and to Professor David D. Yao (Columbia University, formerly of Harvard University), both of whom painstakingly read portions of the original first edition manuscript and offered useful comments and valuable suggestions.

Our eldest son, (Professor) Deepankar Medhi (University of Missouri, Kansas City, formerly of AT&T Bell Laboratories), and our eldest daughter, Shakuntala Choudhury (AT&T Technology Systems, Bridgewater, NJ), rendered invaluable technical assistance. Our younger son, Shubhankar, and younger daughter, Alakanandaa have been of great help. Interest has been shown also by our granddaughters, Namrata Gargee Choudhury (now at University of Pennsylvania) and Sumangala, and grandsons Neiloy, Robby, Abhishek Vikram, and Shimankar.

After the first edition appeared I received feedback from several users and experts. At my specific request, Professor Pavel P. Bocharov, Moscow

(reviewer of the first edition in *Mathematical Reviews*) took great pains to look into the first edition carefully and was kind enough to offer concrete suggestions for improvement. I am immensely grateful to him. I would like to thank Professors Sheldon Ross, Svetlozar Rachev, Donald Miller, Chun Jin, and Morteza Shafii-Mousavi, as well as Dr. Patrick L. Reilly (then with Motorola) for their many helpful comments. I thank Dr. A. Borthakur, Dr. G. Choudhury, and Dr. K. K. Das, who helped me with proofreading. My very special thanks are due to Dr. Das who along with Mitra also managed much of the typesetting in LaTex with great efficiency. Deepankar provided me with interesting material and references. Our grandson Riddhiman was a source of inspiration.

Last but not least, I thank my wife, Prity, who bore with me patiently through the long hours that kept me engaged for months and months and who seldom complained of (or was tired of) waiting!

Assistance from the Department of Science and Technology, Government of India, is gratefully acknowledged. Thanks are also due to Ms. Barbara Holland, Ms. Nancy Zachor, Mr. Tom Singer and other concerned individuals of Academic Press for their care and cooperation.

July 1, 2002 Jyotiprasad Medhi

Stochastic Processes

1.1 Introduction

The theory of stochastic processes mainly originated from the needs of physicists. It began with the study of physical phenomena as random phenomena changing with time. Let t be a parameter assuming values in a set T, and let $X(t)$ represent a random or stochastic variable for every $t \in T$. The family or collection of random variables $\{X(t), t \in T\}$ is called a stochastic process. The parameter or index t is generally interpreted as time and the random variable $X(t)$ as the state of the process at time t. The elements of T are time points, or epochs, and T is a linear set, denumerable or nondenumerable. If T is countable (or denumerable), then the stochastic process $\{X(t), t \in T\}$ is said to be a discrete-parameter or discrete-time process, while if T is an interval of the real line, then the stochastic process is said to be a continuous-parameter or continuous-time process. For example, $\{X_n, n = 0, 1, 2, \ldots\}$ is a discrete-time and $\{X(t), t \geq 0\}$ is a continuous-time process. The set of all possible values that the random variable $X(t)$ can assume is called the state space of the process; this set may be countable or noncountable. Thus, stochastic processes may be classified into these four types:

(i) discrete-time and discrete state space,

(ii) discrete-time and continuous state space,

(iii) continuous-time and discrete state space, and

(iv) continuous-time and continuous state space.

A discrete state space process is often referred to as a chain. A process such as (i) is a discrete-time chain, and a process such as (iii) is a continuous-time chain.

A stochastic process—that is, a family of random variables—thus provides description of the evolution of some physical phenomenon through time. Queueing systems provide many examples of stochastic processes. For example, $X(t)$ might be the number of customers that arrive before a service counter by time t; then $\{X(t), t \geq 0\}$ is of the type (iii) above. Again W_n might be the queueing time of the nth arrival; then $\{W_n, n = 0, 1, 2, \ldots\}$ is of the type (ii) above.

Stochastic processes play an important role in modeling queueing systems. Certain stochastic processes are briefly discussed in this chapter.

1.2 Markov[1] Chains

1.2.1 Basic ideas

Suppose that we observe the state of a system at a discrete set of time points $t = 0, 1, 2, \ldots$. The observations at successive time points define a set of random variables (RVs) $X_0, X_1, X_2, \ldots$. The values assumed by the RVs X_n are the states of the system at time n. Assume that X_n assumes the finite set of values $0, 1, \ldots, m$; then $X_n = i$ implies that the state of the system at time n is i. The family of random variables (RVs) $\{X_n, n \geq 0\}$ is a stochastic process with discrete parameter space $n = 0, 1, 2, \ldots$ and discrete state space $S = \{0, 1, \ldots, m\}$.

Definition 1.1. A stochastic process $\{X_n, n \geq 0\}$ is called a Markov chain, if for every $x_i \in S$,

$$\begin{aligned} &Pr\{X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0\} \\ &\quad = Pr\{X_n = x_n \mid X_{n-1} = x_{n-1}\}, \end{aligned} \tag{1.2.1}$$

provided the first member (LHS) is defined. Equation (1.2.1) indicates a kind of dependence between the RVs X_n; intuitively, it implies that given the present state of the system, the future is independent of the past. The conditional probability

$$Pr\{X_n = k \mid X_{n-1} = j\}, \quad j, k \in S$$

is called the *transition probability* from state j to state k. This is denoted by

$$p_{jk}(n) = Pr\{X_n = k \mid X_{n-1} = j\}. \tag{1.2.2}$$

[1] A. A. Markov (1856–1922)

The Markov chain will be called (temporally) *homogeneous* if $p_{jk}(n)$ does not depend on n—that is,

$$Pr\{X_n = k \mid X_{n-1} = j\} = Pr\{X_{n+m} = k \mid X_{n+m-1} = j\}$$

for $m = -(n-1), -(n-2), \ldots, 0, 1, \ldots$. In such cases we denote $p_{jk}(n)$ simply by p_{jk}. The transition probability p_{jk}, which is the probability of transition from state j to state k in one step—that is, from step $n-1$ to next step n (or from step $n+m-1$ to step $n+m$)—is called one-step transition probability; the transition probability

$$Pr\{X_{r+n} = k \mid X_r = j\}$$

from state j to state k in n steps (from state j in step r to state k in step $r+n$) is called the n-step transition probability. We denote

$$p_{jk}^{(n)} = Pr\{X_{r+n} = k \mid X_r = j\} \tag{1.2.3}$$

so that $p_{jk}^{(1)} = p_{jk}$. Define

$$p_{jk}^{(0)} = \begin{cases} 1, & k = j \\ 0, & k \neq j. \end{cases}$$

Then (1.2.3) is defined for $n = 0, 1, 2, \ldots$. Denote

$$\pi_j = Pr\{X_0 = j\} \quad \text{and} \quad \boldsymbol{\pi}(0) = \{\pi_0, \pi_1, \ldots, \pi_m\};$$

$\boldsymbol{\pi}(0)$ is the initial probability vector. We have

$$\begin{aligned} Pr\{X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n\} &= Pr\{X_0 = x_0, \ldots, X_{n-1} = x_{n-1}\} \\ &\quad \times Pr\{X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}\} \\ &= Pr\{X_0 = x_0\}\, p_{x_0 x_1} p_{x_1 x_2} \cdots p_{x_{n-1} x_n}. \end{aligned} \tag{1.2.4}$$

Thus, given p_{jk} and $\boldsymbol{\pi}(0)$, the joint probability given by (1.2.4) can be determined.

The matrix $\boldsymbol{P} = (p_{jk})$, $j, k \in S$ is called the transition matrix or *transition probability matrix* (TPM) of the Markov chain. $\boldsymbol{P}$ is a nonnegative square matrix with unit row sums—that is, $0 \leq p_{jk} \leq 1, \sum_k p_{jk} = 1$ for every $j \in S$.

A nonnegative square matrix P with unit row sums is called a *stochastic matrix*.

It can be easily shown that P^n is also a stochastic matrix and that

$$\left(p_{jk}^{(n)}\right) = P^n. \tag{1.2.5}$$

That is,

$$p_{jk}^{(2)} = \sum_{r \in s} p_{jr} p_{rk} \quad \text{for every } j,\ k \in S$$

$$p_{jk}^{(n)} = \sum p_{jr}^{(n-1)} p_{rk},$$

more generally,

$$p_{jk}^{(m+n)} = \sum_r p_{jr}^{(m)} p_{rk}^{(n)}$$
$$= \sum_r p_{jr}^{(n)} p_{rk}^{(m)}, \quad r \in S. \tag{1.2.6}$$

Equation (1.2.6) is a special case of the *Chapman-Kolmogorov equation*. It is satisfied by transition probabilities of a Markov chain.

Remark: To every stochastic matrix $P = (p_{ij}), i, j = 0, 1, \ldots$ there exists a homogeneous Markov chain $\{X_n, n = 0, 1, \ldots\}$ with state space $S = \{0, 1, \ldots\}$ and one-step transition probability $p_{ij}, i, j \in S$.

That is, to every stochastic matrix P, there corresponds a Markov chain $\{X_n\}$ for which P is the unit-step transition matrix.

Then $P^2 = (p_{ij}^{(2)})$ is also stochastic; it is the two-step transition matrix for the chain $\{X_n, n = 0, 1, \ldots\}$. However, not every stochastic matrix is the two-step transition matrix of a Markov chain.

1.2.2 Classification of states and chains

1.2.2.1 Finite homogeneous chain

Let $\{X_n, n \geq 0\}$ be a finite homogeneous Markov chain having TPM $P = (p_{jk})$ and state space S, and let i, j, k be arbitrary states of S.

State i is said to lead to state j (or state j is said to be accessible from state i) and is denoted by $i \to j$, if there exists an integer $m(\geq 1)$ such that $p_{ij}^{(m)} > 0$. If no such integer exists, then we say that i does not lead to j: denote this by $i \not\to j$. Two states are said to communicate with each other if $i \to j$ and $j \to i$; this is denoted by $i \leftrightarrow j$. The relations $i \to j$ and $i \leftrightarrow j$ are transitive.

One way to classify the state of a chain is given below.

If $i \to j$ but $j \not\to i$, then index i is said to be inessential. If $i \to j$ implies $i \leftrightarrow j$ for at least one j, then i is said to be essential. All essential states can be grouped into a number of essential classes such that all states belonging to an essential class communicate with one another, but cannot lead to a state outside the class. An essential class is closed—that is, if i, j, k belong to an essential class and l is another state outside the class, then $i \leftrightarrow j \leftrightarrow k$, but $i \not\to l$, $j \not\to l$, $k \not\to l$ (though $l \to i, j$, or k).

Inessential states, if any, can be grouped into a number of inessential classes such that all states belonging to an inessential class communicate with all states in the class. A finite homogeneous Markov chain has at least one essential class of states, whereas a Markov chain with denumerable number of states may not necessarily have any essential class.

A Markov chain is said to be *irreducible* if it contains exactly one essential class of states; in this case every state communicates with every other state of the chain. A nonirreducible (or reducible) chain may have more than one essential class of states as well as some inessential classes.

Suppose that $i \to i$—that is, there exists some $m(\geq 1)$ such that $p_{ii}^{(m)} > 0$. The greatest common divisor of all such m for which $p_{ii}^{(m)} > 0$ is called the period $d(i)$ of the state i. If $d(i) = 1$, then state i is said to be *aperiodic* (or *acyclic*), and if $d(i) > 1$, state i is said to be *periodic* with period $d(i)$. If $p_{ii} > 0$, then clearly state i is aperiodic.

For an irreducible Markov chain, either all states are aperiodic or they are periodic having the same period $d(i) = d(j) = \cdots$. Thus, irreducible Markov chains can be divided into two classes: aperiodic and periodic. An irreducible Markov chain is said to be *primitive* if it is aperiodic and *imprimitive* if it is periodic.

A Markov chain whose essential states form a single essential class and are aperiodic is said to be *regular*. Such a chain may have some inessential indices as well. Note that the transitions between states of the essential class of a regular chain form a submatrix P_1 that is stochastic. The TPM P of a regular chain can be written in canonical form,

$$P = \begin{pmatrix} P_1 & 0 \\ R_1 & Q \end{pmatrix} \tag{1.2.7}$$

where the stochastic submatrix P_1 corresponds to transitions between states of essential class, the square matrix Q to transitions between inessential states, and the rectangular matrix R_1 to transitions between inessential states and essential states (see Seneta (1981) for details).

1.2.2.2 Ergodicity property

We shall now discuss an important concept: the ergodicity property. The classification given above is enough for this discussion so far as finite chains are concerned.

Before considering ergodicity, we shall describe another concept: invariant measure.

If $\{X_n\}$ is a Markov chain with TPM P, and if there exists a probability vector $\boldsymbol{V} = (v_1, v_2, \ldots)$ (i.e., $0 \leq v_i \leq 1, \sum v_i = 1$) such that

$$\boldsymbol{V}P = \boldsymbol{V},$$

then $\boldsymbol{V}$ is called an *invariant measure* (or *stationary distribution*) of the Markov chain $\{X_n\}$ or with respect to the stochastic matrix P.

If there exists $\boldsymbol{V}$ such that $\boldsymbol{V}P \leq \boldsymbol{V}$, then $\boldsymbol{V}$ is called a subinvariant measure of the chain with TPM P. Our interest lies in the types of Markov chains that possess invariant measures.

For a finite, irreducible Markov chain with TPM P, an invariant measure exists and is unique—that is, there is a unique probability vector $\boldsymbol{V}$ such that

$$\boldsymbol{V}P = \boldsymbol{V}, \qquad \boldsymbol{V}\boldsymbol{e} = 1, \tag{1.2.8}$$

where $\boldsymbol{e} = (1, 1, \ldots, 1)$ is a column vector with all its elements equal to unity. We next discuss the limiting behavior of chains.

1.2.2.3 Ergodic theorems

Theorem 1.1. Ergodic Theorem for Primitive Chains
Let $\{X_n, n \geq 0\}$ be a finite irreducible aperiodic Markov chain with TPM P and state space S. Then, as $n \to \infty$,

$$P^n \to \boldsymbol{e}\boldsymbol{V} \tag{1.2.9}$$

elementwise, where $\boldsymbol{V}$ is the unique stationary distribution (or invariant measure) of the chain.

Further, the rate of approach to the limit is geometrically fast—that is, there exist positive constants $a, b, 0 < b < 1$ such that $\varepsilon_{ij}^{(n)} \leq ab^n$, where $p_{ij}^{(n)} = v_j + \varepsilon_{ij}^{(n)}$.

The above theorem implies that, for every $j \in S$,

$$\lim_{n\to\infty} p_{ij}^{(n)} \to v_j \tag{1.2.10}$$

exists and is independent of the initial state i and the quantities v_j's are given by the solution of the matrix equation

$$\boldsymbol{V}P = \boldsymbol{V} \quad with\ \boldsymbol{V}\boldsymbol{e} = 1. \tag{1.2.11}$$

The limiting probability distribution of the chain tends to an equilibrium distribution that is independent of the initial distribution. This tendency is known as ergodicity or ergodic property.

Another ergodic theorem is stated below.

Theorem 1.2. *Let $\{X_n, n \geq 0\}$ be a finite k state regular Markov chain (i.e., having a single essential class of aperiodic states) and having TPM P. Let $\boldsymbol{V}_1$ be the stationary distribution corresponding to the primitive submatrix P_1 (corresponding to the transitions between the states of the essential aperiodic class).*

Let $\boldsymbol{V} = (V_1, 0)$ be a $1 \times k$ vector. Then, as $n \to \infty$,

$$P^n \to \boldsymbol{e}\boldsymbol{V} \tag{1.2.9a}$$

elementwise.

$\boldsymbol{V}$ is the unique stationary distribution corresponding to the matrix P and the rate of approach to the limit in (1.2.9a) is geometrically fast.

If C denotes the single essential class with states j, $j = 1, 2, \ldots, m (m < k)$, $j \in C$, and $\boldsymbol{V}_1 = (v_1, v_2, \ldots, v_j, \ldots, v_m)$ is given by the solution of

$$\boldsymbol{V}_1 P_1 = \boldsymbol{V}_1, \qquad \boldsymbol{V}_1\boldsymbol{e} = 1, \tag{1.2.12}$$

then the above result implies that, as $n \to \infty$,

$$\begin{aligned} &\lim p_{ij}^{(n)} \to v_j, \qquad j \in C, \quad \text{and} \\ &\lim p_{ij}^{(n)} \to 0, \qquad j \notin C. \end{aligned} \tag{1.2.13}$$

The above theorem asserts that regularity of the chain is a sufficient condition for ergodicity. It is also a necessary condition.

Example 1.1. Consider the Markov chain having state space $S = \{0, 1, 2\}$ and TPM P

$$P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} & 0 \end{pmatrix}$$

This is a finite irreducible chain. Its invariant measure $\boldsymbol{V} = (v_1, v_2, v_3)$ is given by the solution of

$$\boldsymbol{V}P = \boldsymbol{V},$$

which leads to

$$\begin{aligned} v_1 &= \frac{1}{2}v_2 + \frac{3}{4}v_3 \\ v_2 &= \frac{1}{3}v_1 + \frac{1}{4}v_3 \\ v_3 &= \frac{2}{3}v_1 + \frac{1}{2}v_2. \end{aligned}$$

As $\boldsymbol{V}\boldsymbol{e} = v_1 + v_2 + v_3 = 1$, one of the above equations is redundant. We get

$$\boldsymbol{V} = (v_1, v_2, v_3),$$

where

$$v_1 = \frac{21}{53}, \quad v_2 = \frac{12}{53}, \quad v_3 = \frac{20}{53}.$$

Thus,

$$P^n \to \mathbf{e}\boldsymbol{V}.$$

That is, as $n \to \infty$, for all $i = 1, 2, 3$,

$$\lim p_{i1}^{(n)} = \frac{21}{53}, \qquad \lim p_{i2}^{(n)} = \frac{12}{53}, \qquad \lim p_{i3}^{(n)} = \frac{20}{53}.$$

Example 1.2. Consider the two-state Markov chain with TPM

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \quad 0 \le p \le 1.$$

The equation $VP = V, (V = (v_1, v_2))$ leads to

$$v_1 = v_2 = \frac{1}{2}, \quad \text{for all } p,$$

so that the invariant distribution is $(\frac{1}{2}, \frac{1}{2})$ for all p. We have, for $p \neq 0$

$$P^n = \boldsymbol{e}\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}(1-2p)^n \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

For $p = 0$, the chain consists of two absorbing states 0 and 1—that is, it consists of two essential classes, C_1 and C_2, with members 0 and 1, respectively. The chain is decomposable. For $p = 1$,

$$P^n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ when } n \text{ is even and}$$

$$= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ when } n \text{ is odd.}$$

The chain is periodic.

Thus, even though invariant distribution exists for all $p, 0 \leq p \leq 1$, as $n \to \infty$,

$$\lim p_{i0}^{(n)} = \frac{1}{2}$$

$$\text{and} \quad \lim p_{i1}^{(n)} = \frac{1}{2}$$

exist only when $p \neq 0, 1$.

Remarks: We can make two deductions: (1) Existence of stationary distribution (i.e., existence of a solution of $\boldsymbol{VP} = \boldsymbol{V}, \boldsymbol{Ve} = 1$) does not necessarily imply existence of the limiting distribution

$$\left\{\lim_{n\to\infty} p_{ij}^{(n)}\right\}$$

(2) The example also shows how much the transient behavior can vary over the class of transient matrices for a given equilibrium distribution.

See Whitt (1983) for a discussion on this topic.

1.2.2.4 Markov chain having a denumerably infinite number of states

So far we have considered homogeneous Markov chains with a finite number of states. Now we shall discuss homogeneous chains having a denumerably infinite number of states. We shall denote the state space by $S = \{0, 1, 2, \ldots\}$ instead of the more general

$$S = \{\ldots, -3, -2, -1, 0, 1, 2, \ldots\}$$

Notions already defined (e.g., accessibility, communication, and periodicity) and subsequent definitions of essential and inessential states, essential classes, inessential classes, irreducible chains, and primitive chains will remain valid for a chain with a denumerable number of states. Only the definition of a regular chain cannot be carried over to this case. For whereas in a finite chain there is at least one essential class (and so the states of a finite chain may constitute exactly one essential class or more than one essential class), there may not be any essential class in the case of a chain with a denumerably infinite number of states. For example, consider the chain $\{X_n, n \geq 0\}$ with $S = \{0, 1, 2, \ldots\}$ and TPM $P = (p_{ij})$ where

$$\begin{aligned} p_{ij} &= 1 \quad j = i + 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

This denumerable chain does not possess any essential class at all. Each state is inessential, and no two states communicate. Whereas consideration of classification of states into essential and inessential classes was adequate for dealing with limiting distribution for finite chains, a more sensitive classification of states would be required in the present case of chains with a denumerable infinity of states.

1.2.2.5 Transience and Recurrence

Define

$$\begin{aligned} &\{f_{ij}^{(n)}\}, i, j = 1, 2, \ldots, n, \\ &\qquad f_{ij}^{(0)} = 0, \qquad f_{ij}^{(1)} = p_{ij}, \text{ and} \\ &f_{ij}^{(k+1)} = \sum_{\substack{r \\ r \neq j}} p_{ir} f_{rj}^{(k)}, \qquad k \geq 1. \end{aligned} \tag{1.2.14}$$

The quantity $f_{ij}^{(k)}$ is the probability of transition from state i to state j in k steps, without revisiting the state j in the meantime. (It is called *taboo* probability—j being the *taboo* state.) Here $\{f_{ij}^{(k)}\}$ gives the distribution of the first passage time from state i to state j. We can write

$$f_{ij}^{(n)} = Pr\{X_n = j, X_r \neq j, \ r = 1, 2, \ldots, n-1 \mid X_0 = i\}$$

The relation (1.2.14) can also be written as

$$\begin{aligned} p_{ij}^{(n)} &= \sum_{r=0}^{n} f_{ij}^{(r)} p_{jj}^{(n-r)}, \quad n \geq 1 \\ &= \sum_{r=0}^{n} f_{ij}^{(n-r)} p_{jj}^{(r)}. \end{aligned} \tag{1.2.15}$$

The relations (1.2.14) and (1.2.15) are known as *first entrance formulas.*

Let

$$P_{ij}(s) = \sum_n p_{ij}^{(n)} s^n, \qquad F_{ij}(s) = \sum_n f_{ij}^{(n)} s^n, \quad |s| < 1.$$

Then from the convolution structure,

$$\begin{aligned} P_{ij}(s) &= P_{ii}(s) F_{ij}(s), \quad j \neq i \\ P_{ii}(s) - 1 &= P_{ii}(s) F_{ii}(s). \end{aligned} \tag{1.2.16}$$

Definition 1.2. A state i is said to be *persistent* if $F_{ii} = F_{ii}(1-0) = 1$ and is said to be *transient* if $F_{ii}(1-0) < 1$.

A persistent state is null or nonnull based on whether $\mu_{ii} = F_{ii}'(1) = \infty$ or $< \infty$, respectively.

Equivalent criteria of persistence and recurrence are as follows.

An index i is persistent *iff* (if and only if)

$$\sum_n p_{ii}^{(n)} = \infty$$

and is transient *iff*

$$\sum_n p_{ii}^{(n)} < \infty.$$

The relationship between these two types of classification of states and chain can be given as follows.

An inessential state is transient and a persistent state is essential. In the case of a finite chain, i is transient *iff* it is inessential; otherwise it is nonnull persistent.

All the states of an irreducible chain, whether finite or denumerable, are of the same type: all transient, all null persistent, or all nonnull persistent.

A finite Markov chain contains at least one persistent state. Further, a finite irreducible Markov chain is nonnull persistent. The ergodic theorem for a Markov chain with a denumerable infinite number of states is stated below.

Theorem 1.3. General Ergodic Theorem

Let P be the TPM of an irreducible aperiodic (i.e., primitive) Markov chain with a countable state space S (which may have a finite or a denumerably infinite number of states). If the Markov chain is transient or null persistent, then for each $i, j \in S$,

$$\lim_{n \to \infty} p_{ij}^{(n)} \to 0. \tag{1.2.17a}$$

If the chain is nonnull persistent, then for each $i, j \in S$,

$$\lim_{n \to \infty} p_{ij}^{(n)} = v_j \tag{1.2.17b}$$

exists and is independent of i. The probability vector $V = (v_1, v_2, \ldots)$ *is the unique invariant measure of P—that is,*

$$VP = V, \qquad Ve = 1; \tag{1.2.18a}$$

and further if μ_{jj} *is the mean recurrence time of state j, then*

$$v_j = (\mu_{jj})^{-1}. \tag{1.2.18b}$$

The result is general and holds for a chain with a countable state space S. In case the chain is finite, irreducibility ensures nonnull persistence, so that irreducibility and aperiodicity (i.e., primitivity) constitute a set of sufficient conditions for ergodicity of a finite chain. The sufficient conditions for ergodicity ($\lim p_{ij}^{(n)} = v_i$) *for a chain with a denumerably infinite number of states involve, besides irreducibility and aperiodicity, nonnull persistence of the chain. For a chain with a denumerably infinite number of states, the number of equations given by (1.2.18a) will be infinite. It would sometimes be more convenient to find* V *in terms of the generating function of* $\{v_j\}$ *than to attempt to solve Eqn. (1.2.18a) as such. We shall consider two such Markov chains that arise in queueing theory. See the Note below.*

Example 1.3. Consider a Markov chain with state space $S = \{0, 1, 2, \ldots\}$ having a denumerable number of states and having TPM

$$P = \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & p_0 & p_1 & p_2 & \cdots \\ 0 & 0 & p_0 & p_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{1.2.19}$$

where $\sum_k p_k = 1$. Let

$$P(s) = \sum_k p_k s^k \quad \text{and} \quad V(s) = \sum_k v_k s^k, \quad |s| < 1$$

be the probability-generating functions (PGF) of $\{p_k\}$ and $\{v_k\}$, respectively. Clearly, the chain is irreducible and aperiodic; since it is a denumerable chain, we need to consider transience and persistence of the chain to study its ergodic property.

It can be shown that the states of the chain (which are all of the same type because of the irreducibility of the chain) are transient, persistent null or persistent nonnull according to

$$P'(1) > 1, \qquad P'(1) = 1, \qquad P'(1) < 1,$$

respectively (see Prabhu, 1965). Assume that $P'(1) < 1$, so that the states are persistent nonnull; then from (1.2.18a), we get

$$v_k = p_k v_0 + p_k v_1 + p_{k-1} v_2 + \cdots + p_0 v_{k+1}, \quad k \geq 0. \tag{1.2.20}$$

Multiplying both sides of (1.2.20) by s^k and adding over $k = 0, 1, 2, \ldots$, we get

$$\begin{aligned} V(s) &= v_0 P(s) + v_1 P(s) + v_2 s P(s) + \cdots + v_{k+1} s^k P(s) + \cdots \\ &= P(s)[v_0 + (V(s) - v_0)/s]; \end{aligned}$$

whence

$$V(s) = \frac{v_0(1-s)P(s)}{P(s) - s}.$$

Since $V(1) = 1$, we have

$$\begin{aligned} 1 = \lim_{s\to 1} V(s) &= v_0 \left[\lim_{s\to 1} \frac{(1-s)P(s)}{P(s)-s}\right] \\ &= v_0 \left[\frac{1}{1 - P'(1)}\right]. \end{aligned}$$

Thus,

$$V(s) = \frac{(1 - P'(1))(1-s)P(s)}{P(s) - s}. \tag{1.2.21}$$

Example 1.4. Consider a Markov chain with state space $S = \{0, 1, 2, \ldots\}$ and having TPM

$$P = \begin{bmatrix} h_0 & g_0 & 0 & 0 & 0 & \cdots \\ h_1 & g_1 & g_0 & 0 & 0 & \cdots \\ h_2 & g_2 & g_1 & g_0 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{1.2.22}$$

where $h_i = g_{i+1} + g_{i+2} + \cdots, i \geq 0, g_i > 0$, and $\sum_{i=0} g_i = 1$. Here $p_{i0} = h_i, i \geq 0$,

$$\begin{aligned} p_{ij} &= g_{i+1-j}, \qquad i + 1 \geq j \geq 1, \quad i \geq 0 \\ &= 0, \qquad i + 1 < j. \end{aligned}$$

The chain is irreducible and aperiodic. It can be shown that it is persistent non-null when $\alpha = \sum j g_j > 1$. Then the chain is ergodic and $v_j = \lim_{n\to\infty} p_{ij}^{(n)}$ exist and are given as a solution of (1.2.18a); these lead to

$$v_0 = \sum v_r h_r \tag{1.2.23a}$$

$$v_j = \sum v_{r+j-1} g_r, \quad j \geq 1 \tag{1.2.23b}$$

$$\sum_{j=0} v_j = 1 \tag{1.2.23c}$$

Let $G(s) = \sum_r g_r s^r$ be the PGF of $\{g_r\}$. Denote the displacement operator by E so that

$$E^r(v_k) = v_{k+r}, \quad r = 0, 1, 2, \ldots$$

Then we can write (1.2.23b) in symbols as

$$E(v_{j-1}) = v_j = \sum g_r E^r(v_{j-1}), \quad j \geq 1 \quad \text{or}$$

$$\left\{E - \sum g_r E^r\right\} v_j = 0, \quad j \geq 0 \quad \text{or} \tag{1.2.24}$$

$$\{E - G(E)\}v_j = 0, \quad j \geq 0.$$

The characteristic equation of the above difference equation is given by

$$r(z) \equiv z - G(z) = 0. \tag{1.2.25}$$

It can be shown that when $\alpha = G'(1) > 1$, there is exactly one real root of $r(z) = 0$ between 0 and 1. Denote this root by r_0 and the other roots by $r_1, r_2, \ldots, |r_i| > 1, i \geq 1$. The solution of (1.2.24) can thus be put as

$$v_j = c_0 r_0^j + \sum_{i=j} c_i r_i^j, \quad j \geq 0,$$

where the c's are constants. Since $\sum v_j = 1$,

$$c_i \equiv 0 \quad \text{for } i \geq 1,$$
$$v_j = c_0 r_0^j, \quad j \geq 0, \quad \text{and}$$
$$c_0 = 1 - r_0$$

so that

$$v_j = (1 - r_0) r_0^j, \quad j \geq 0, \tag{1.2.26}$$

r_0 being the root lying between 0 and 1 of (1.2.25) (provided $\alpha = G'(1) > 1$). The distribution is geometric.

Notes:

(1) The equation $VP = V$ is quite well known in the matrix theory. It follows from the well-known Perron-Frobenius theorem of matrix theory that there exists a solution $V = (v_1, v_2, \ldots)$ of the matrix equation $VP = V$ subject to the constraints $v_i \geq 0, \sum v_i = 1$.

(2) When the order of P is not large, the equations can be solved fairly easily to get $V = (v_1, v_2, \ldots)$. When the order of P is large (infinite), the number of equations is also large (infinite) and the solution of the equations becomes troublesome. In Example 1.3 we considered and obtained the solution in terms of the generating function $V(s) = \sum v_j s^j$. This method may not always be applicable.

See also Remarks (4) in Section 1.3.5.

1.3 Continuous-Time Markov Chains

We shall now consider continuous-time Markov chains—that is, Markov processes with discrete state space. Let $\{X(t), 0 \leq t < \infty\}$ be a Markov process with countable state space $S = \{0, 1, 2, \ldots\}$. We assume that the chain is temporally homogeneous. The transition probability function given by

$$p_{ij}(t) = Pr\{X(t+u) = j \mid X(u) = i\}, \quad t > 0, \quad i, j \in S, \tag{1.3.1}$$

is then independent of $u \geq 0$. We have for all t,

$$0 \leq p_{ij}(t) \leq 1, \quad \sum_j p_{ij}(t) = 1, \quad \text{for all } j \in S.$$

Denote the matrix of transition probabilities by

$$P(t) = (p_{ij}(t)), \quad i, j \in S.$$

Setting $p_{ij}(0) = \delta_{ij}$, the initial condition can be put as

$$\boldsymbol{P}(0) = \boldsymbol{I}.$$

Denote the probability that the system is at state j at time t by

$$\pi_j(t) = Pr\{X(t) = j\};$$

the vector $\pi(t) = \{\pi_1(t), \pi_2(t), \ldots\}$ is the probability vector of the state of the system at time t; $\pi(0)$ is the initial probability vector. Now

$$\begin{aligned} \pi_j(t) &= \sum_i Pr\{X(t+u) = j \mid X(u) = i\} Pr\{X(u) = i\} \\ &= \sum_i p_{ij}(t)\ Pr\{X(0) = i\} \qquad (1.3.2) \\ &= \sum_i p_{ij}(t)\pi_i(0). \end{aligned}$$

Thus, given initial probability vector $\pi(0)$ and the transition functions $p_{ij}(t)$, the state probabilities can be calculated and the probabilistic behavior of the system can be completely determined. The matrix form of (1.3.2) is

$$\pi(t) = \pi(0)P(t). \tag{1.3.3}$$

1.3.1 Sojourn time

The time taken (or the waiting time) for change of state from state i is a random variable—say, τ_i; that is, the sojourn time at state i is τ_i. Then

$$\begin{aligned} &Pr\{\tau_i > s + t \mid X(0) = i\} \\ &\quad = Pr\{\tau_i > s + t \mid X(0) = i,\ \tau_i > s\} \\ &\qquad \times Pr\{\tau_i > s \mid X(0) = i\}, \quad t \geq 0. \end{aligned} \tag{1.3.4}$$

Denote

$$\bar{F}_i(u) = Pr\{\tau_i > u \mid X(0) = i\}, \quad u \geq 0.$$

Then (1.3.4) can be written as

$$\bar{F}_i(t+s) = \bar{F}_i(t)\bar{F}_i(s), \quad s, t \geq 0;$$

$\bar{F}(.)$ is right continuous; the only right continuous solution of the functional equation is

$$\bar{F}_i(u) = e^{-a_i u}, \quad u \geq 0, \quad a_i > 0 \text{ is a constant.} \tag{1.3.5}$$

That is, sojourn time τ_i at state i is exponential with parameter a_i. Further, the sojourn times τ_i and τ_j are independent. We have, for $t \geq 0$, $T \geq 0$,

$$p_{ij}(T+t) = \sum_k p_{ik}(T)\, p_{kj}(t), \quad i, j, k \in S \tag{1.3.6}$$

or, in matrix form,

$$P(T+t) = P(T)P(t), \tag{1.3.7}$$

which is called the *Chapman-Kolmogorov equation.*

1.3.2 Transition density matrix or infinitesimal generator

Denote the right-hand derivative at $t = 0$, by

$$\begin{aligned} q_{ij} &= \lim_{h\to 0} \frac{p_{ij}(h) - p_{ij}(0)}{h} = \lim_{h\to 0} \frac{p_{ij}(h)}{h}, \quad i \neq j \quad \text{and} \\ q_{ii} &= \lim_{h\to 0} \frac{p_{ii}(h) - p_{ii}(0)}{h} = \lim_{h\to 0} \frac{p_{ii}(h) - 1}{h}; \end{aligned} \tag{1.3.8}$$

write $-q_{ii} = q_i$. It is to be noted that $q_{ij}, i \neq j$ is always finite. While $q_i(\geq 0)$ always exists and is finite when S is finite, q_i may be infinite when S is denumerably infinite. Writing $\boldsymbol{Q} = (q_{ij})$, we can denote (1.3.8) in matrix notation as

$$Q = \lim_{h\to 0} \frac{P(h) - \boldsymbol{I}}{h}.$$

From (1.3.8) it follows that, for small h,

$$\begin{aligned} p_{ij}(h) &= hq_{ij} + o(h), \quad i \neq j, \\ p_{ii}(h) &= hq_i + o(h), \end{aligned} \tag{1.3.9}$$

where $o(h)$ is used as a symbol to denote a function of h that tends to zero more rapidly than h; that is, $o(h)/h \to 0$ as $h \to 0$. Again,

$$\sum_j p_{ij}(h) = 1, \quad \text{or}$$

$$\sum_{j \neq i} p_{ij}(h) + p_{ii}(h) - 1 = 0; \quad \text{whence we get}$$

$$\sum_{j \neq i} q_{ij} + q_{ii} = 0$$

$$\text{or} \qquad \sum_{j \neq i} q_{ij} = q_i. \tag{1.3.10}$$

The matrix $Q = (q_{ij})$ is called the *transition density matrix* or *infinitesimal generator* or *rate matrix* or simply Q-matrix. The Q-matrix is such that (i) its diagonal elements are negative and off-diagonal elements are positive, and (ii) each row sum is zero. Let $S = \{0, 1, 2, \ldots, m\}$ be a finite set, then

$$Q = \begin{bmatrix} -q_0 & q_{01} & \cdots & q_{0m} \\ q_{10} & -q_1 & \cdots & q_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ q_{m0} & q_{m1} & \cdots & -q_m \end{bmatrix}$$

1.3.3 Limiting behavior: ergodicity

The states of a continuous-time Markov chain admit of a classification similar to those of a discrete-time chain. A state j is said to be *accessible* or *reachable* from state $i (i \to j)$ if, for some $t > 0$, $p_{ij}(t) > 0$. States i and j communicate if $i \to j$ and $j \to i$. A continuous-time Markov chain is said to be *irreducible* if every state can be reached from every other state (or if each pair of states communicates).

Let α_{ij} denote the (first) entrance time from state i to state j without visiting j in the meantime and let $F(.)$ denote its DF—that is,

$$F_{ij}(t) = Pr\{\alpha_{ij} < t\}, \quad t > 0,$$
$$= 0, \quad t \leq 0.$$

A state i is called *persistent* if

$$\lim_{t \to \infty} F_{ii}(t) = 1$$

and *transient* otherwise.

Criteria of transience and persistence can be expressed in terms of $p_{ij}(t)$ as follows:

State i is transient *iff*

$$\int_0^\infty p_{ii}(t)dt < \infty.$$

If state i, is *null-persistent*, then

$$\lim_{t\to\infty} p_{ii}(t) = 0,$$

and if state i is nonnull-persistent, then

$$\lim_{t\to\infty} p_{ii}(t) > 0.$$

From the Chapman-Kolmogorov equation (1.3.7) we get

$$\begin{aligned} p_{ij}(h+t) &= \sum_k p_{ik}(h)\, p_{kj}(t) \\ &= \sum_{k\neq i} p_{ik}(h)\, p_{kj}(t) + p_{ii}(h)\, p_{ij}(t) \end{aligned}$$

so that

$$\frac{p_{ij}(h+t) - p_{ij}(t)}{h} = \sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) + \left(\frac{p_{ii}(h)-1}{h}\right) p_{ij}(t).$$

Taking the limit as $h \to 0$ and assuming that the order of the operations of taking the limit and summation can be interchanged, we get

$$\begin{aligned} \lim_{h\to 0} \frac{p_{ij}(h+t) - p_{ij}(t)}{h} &= \sum_{k\neq i} \left[\lim_{h\to 0} \frac{p_{ik}(h)}{h}\right] p_{kj}(t) \\ &\quad + \left[\lim_{h\to 0} \frac{p_{ii}(h)-1}{h}\right] p_{ik}(t) \end{aligned}$$

$$\text{or} \qquad p'_{ij}(t) = \sum_{k\neq i} q_{ik}\, p_{kj}(t) + q_i\, p_{ij}(t), \tag{1.3.11}$$

which is another form of the Chapman-Kolmogorov (backward) equation; it is in terms of the elements of the Q-matrix. In matrix notation, we get

$$P'(t) = QP(t). \tag{1.3.11a}$$

Again, from (1.3.7) we get

$$\begin{aligned} p_{ij}(t+h) &= \sum_k p_{ik}(t)\, p_{kj}(h) \\ &= \sum_{k\neq j} p_{ik}(t)\, p_{kj}(h) + p_{ij}(t)\, p_{ij}(h). \end{aligned}$$

Assuming that the operations of limit and summation are interchangeable and proceeding as above, we get

$$p'_{ij}(t) = \sum_{k \neq i} p_{ik}(t) q_{kj} + q_j p_{ij}(t), \tag{1.3.12}$$

which is the Chapman-Kolmogorov *forward* equation. In matrix notation,

$$P'(t) = P(t)\,Q. \tag{1.3.12a}$$

Using (1.3.3), we can also put (1.3.11a) and (1.3.12a) in the form

$$\frac{d}{dt}\{\pi(t)\} = Q\pi(t) = \pi(t)\,Q. \tag{1.3.13}$$

1.3.4 Transient solution

Consider a finite $(m+1)$ state chain, with given rate matrix Q. Solving (1.3.11a) or (1.3.12a), we get $\boldsymbol{P}(t) = \boldsymbol{P}(0)e^{Q(t)}$, with $P(0) = I$; we have

$$\boldsymbol{P}(t) = e^{Qt} = I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!} \tag{1.3.14}$$

$$\text{or} \qquad \pi(t) = \pi(0)\left(I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!}\right). \tag{1.3.15}$$

Assume that the eigenvalues d_i of Q are all distinct, $d_i \neq d_j, i, j = 0, 1, \ldots, m$. Let D be the diagonal matrix having $d_0, d_1, \ldots, d_m$ as its diagonal elements. Then there exists a nonsingular matrix H (whose column vectors are right eigenvectors of Q) such that Q can be written in the canonical form

$$Q = HDH^{-1}.$$

Then

$$Q^n = HD^nH^{-1}$$

and, substituting in (1.3.14), we get

$$P(t) = H\Lambda(t)H^{-1} \quad \text{and}$$
$$\pi(t) = \pi(0)P(t), \tag{1.3.16}$$

where $\Lambda(t)$ is the diagonal matrix with diagonal elements $e^{d_i t}$, $i = 0, 1, \ldots, m$.

It may be noted that in the general case when the eigenvalues of the matrix Q are not necessarily distinct, Q can still be expressed in the canonical form $Q = SZS^{-1}$ and $P(t)$ can be obtained as above.

The transient solution can be obtained as given above. While an analytical solution can be obtained, especially when m is small, it becomes difficult when m is large. For such cases, numerical methods have been put forward (Grassman, 1977; Gross and Miller, 1984a,b). See Section 1.6.3.

For many stochastic systems such as queueing systems and reliability systems, computation of the vector $\pi(t)$ transient probabilities is useful. It is especially important when convergence to steady state is slow.

1.3.5 Alternative definition

A continuous-time Markov chain with state space $S = \{0, 1, 2, \ldots\}$ can be defined in another way as follows (Ross, 1980). It is a stochastic process such that (i) each time it enters state i, the time it spends in that state before making a transition to another state $j(\neq i) \in S$—that is, sojourn time in state i is an exponential RV with mean $1/a_i(a_i$ depends on i but not on $j)$; and (ii) when the process leaves state i, it enters another state $j(\neq i)$, with some probability say, p_{ij} (which depends on both i and j), such that, for all i.

$$p_{ii} = 0, \quad 0 \leq p_{ij} \leq 1$$
$$\sum_j p_{ij} = 1, \quad j \in S.$$

Thus, a continuous-time Markov chain is a stochastic process such that (i), its transition from one state to another state of the state space S, is as in a discrete-time Markov chain and (ii) the sojourn in a state i (holding time in state i before moving to another state) is an exponential RV whose parameter depends on i but not on the state next visited. The sojourn times in different states must be independent random variables with exponential distribution.

1.3.5.1 Relationship between p_{ij} and $p_{ij}(t)$

We have

$$p_{ij}(h) = ha_i p_{ij} + o(h),$$

since $p_{ij}(h)$ is the probability that the state of the process changes from i to j in an infinitesimal interval h. Thus,

$$\lim_{h \to o} \frac{p_{ij}(h)}{h} = a_i p_{ij},$$

but by definition LHS equals q_{ij}, so that

$$q_{ij} = a_i p_{ij}. \tag{1.3.17}$$

Again, $1 - p_{ii}(h)$ is the probability that the state of the system changes from state i to some other state in the interval h, so that

$$1 - p_{ii}(h) = a_i h \sum_j p_{ij} + o(h)$$
$$= a_i h + o(h).$$

Thus,

$$\lim_{h\to o} \frac{1 - p_{ii}(h)}{h} = a_i;$$

but by definition LHS equals q_i so that

$$a_i = q_i. \tag{1.3.18}$$

Thus, the Q matrix can also be written as

$$Q = \begin{bmatrix} -a_0 & a_o p_{01} & \cdots & a_0 p_{0m} \\ a_1 p_{10} & -a_1 & \cdots & a_1 p_{1m} \\ \cdots & \cdots & \cdots & \\ a_m p_{m0} & a_m p_{m1} & \cdots & -a_m \end{bmatrix} \tag{1.3.19}$$

We have the corresponding ergodicity property.

Theorem 1.4. Ergodic Theorem
If a Markov chain $\{X(t), t \in T\}$ *is irreducible, then all the states are of the same type.*

In case they are all transient or null-persistent, then

$$\lim_{i\to\infty} p_{ij}(t) = 0, \quad i, j \in S.$$

In case they are nonnull persistent, then

$$\lim_{i\to\infty} p_{ij}(t) = u_j \tag{1.3.20}$$

exists and is independent of the initial state i. Further, $\boldsymbol{U} = (u_1, u_2, \ldots), (\boldsymbol{Ue} = 1)$ *is a probability distribution and is given by the solution of*

$$q_j u_j + \sum_{i\neq j} u_i q_{ij} = 0, \quad i, j \in S, \quad or$$

$$\sum_i u_i q_{ij} = 0, \quad or \tag{1.3.21}$$

$$\boldsymbol{UQ} = \boldsymbol{0}, \quad \boldsymbol{Ue} = 1. \tag{1.3.22}$$

We now consider the alternative definition of the continuous time chain. Using (1.3.17) and (1.3.19), we get from (1.3.21)

$$a_j u_j = \sum_{i\neq j} a_i p_{ij} u_i, \quad j \in S, \quad with \quad \boldsymbol{Ue} = 1, \tag{1.3.23}$$

from which $\boldsymbol{U} = (u_1, u_2, \ldots)$ *can be obtained.*

Remarks

(1) If an irreducible chain is finite, then

$$\lim_{t\to\infty} p_{ij}(t) = u_j, \quad i, j \in S$$

exists. If the chain has a denumerable state space and if it is nonnull persistent, then u_j exists.

(2) When u_j exists, it can be interpreted as the long-run proportion of time the system is in state j.

(3) Equations (1.3.23) have an interesting interpretation.

When the process is in state j, it leaves that state at rate a_j, and u_j is the long-run proportion of time it is in state j, so that $a_j u_j$ = rate at which the process *leaves* state j. Again, when the process is in state i, the rate of transition into state j is $a_i p_{ij} = q_{ij}$, so that

$$\sum_{i\neq j} a_i p_{ij} u_i = \text{ rate at which the process } \textit{enters} \text{ state } j.$$

Thus, Eqn. (1.3.23) can be interpreted as follows: In the long run, the two rates are equal—that is, the rate at which the process enters a state j equals the rate at which it leaves the state j (for each $j \in S$). As the two rates balance each other for every state, Eqn. (1.3.23) are also known as *balance equations.*

The balance equations, as interpreted above, have very useful applications in queueing systems in particular and in stochastic systems in general.

(4) The equations $\boldsymbol{V}\boldsymbol{P} = \boldsymbol{P}$, with $\sum v_i = 1$, (1.2.12) and (1.2.18a) for discrete-time Markov chains written as $\boldsymbol{V}\boldsymbol{Q} = \boldsymbol{O}$, where $\boldsymbol{Q} = \boldsymbol{P} - \boldsymbol{I}$ is a matrix with zero row sums (with $\boldsymbol{V}\boldsymbol{e} = 1$). Then the equations have the same form as in case of continuous-time Markov chains (see Eqn. (1.3.22)). For methods of numerical solutions of Markov Chains, see Stewart (1994). In queueing context (where such equations occur) besides matrix methods (Kaufman, 1983), the graph-theoretic approach (for $\boldsymbol{Q}$ having certain structural properties) has recently been put forward (Tang and Yeung (1999), who also point out the merit and limitation of their approach).

Example 1.5. Two-State Process

Suppose that a system can be in two states: operating and nonoperating or under repair (denoted by 0 and 1, respectively). Suppose that the lengths of the operating and nonoperating periods are independent exponential RVs with parameters a and b, respectively. Let $X(t)$ be the state of the process at time t. $\{X(t), t \geq 0\}$ is a Markov process with state spaces $S = \{0, 1, \}$.

We have, because of exponential distribution,

$$\begin{aligned} p_{01}(h) &= Pr\{\text{change of state from operating to} \\ &\qquad \text{nonoperating in an infinitesimal interval } h\} \\ &= ah + o(h), \quad \text{and so} \\ p_{00}(h) &= 1 - ah + o(h) \\ p_{10}(h) &= bh + o(h), \text{ and} \\ p_{11}(h) &= 1 - bh + o(h) \end{aligned}$$

so that the $\boldsymbol{Q}$-matrix is

$$\boldsymbol{Q} = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}. \tag{1.3.24}$$

The Chapman-Kolmogorov forward equation $\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q}$ gives, for $i = 0, 1$.

$$\begin{aligned} p'_{i0}(t) &= -ap_{i0}(t) + bp_{i1}(t) \\ p'_{i1}(t) &= ap_{i0} - bp_{i1}(t). \end{aligned}$$

Again,

$$p_{i0}(t) = 1 - p_{i1}(t)$$

Assume that $p_{00}(0) = 1$. Solving, we get

$$\begin{aligned} p_{00}(t) &= \frac{b}{a+b} + \frac{a}{a+b}e^{-(a+b)t}, \\ p_{01}(t) &= \frac{a}{a+b} - \frac{a}{a+b}e^{-(a+b)t}, \\ p_{11}(t) &= \frac{a}{a+b} + \frac{b}{a+b}e^{-(a+b)t}, \\ \text{and} \quad p_{10}(t) &= \frac{b}{a+b} - \frac{b}{a+b}e^{-(a+b)t}. \end{aligned}$$

As $t \to \infty$,

$$\begin{aligned} p_{00}(t) &\to \frac{b}{a+b}, \quad p_{01}(t) \to \frac{a}{a+b} \quad \text{and} \\ p_{10}(t) &\to \frac{b}{a+b}, \quad p_{11}(t) \to \frac{a}{a+b}. \end{aligned}$$

These limiting probabilities can also be obtained by using (1.3.21). We have, as $t \to \infty, \lim p_{ij}(t) = u_j$; then from (1.3.21) we get

$$\begin{aligned} q_0 u_0 = u_1 q_{10} &\Rightarrow au_0 = bu_1 \\ &\Rightarrow u_1 = \frac{a}{b}u_0 \end{aligned}$$

and since $u_0 + u_1 = 1$,

$$\lim_{t\to\infty} p_{i0}(t) = u_0 = \frac{b}{a+b}, \quad i = 0, 1$$

and $$\lim_{t\to\infty} p_{i1}(t) = \frac{a}{a+b}, \quad i = 0, 1.$$

1.4 Birth-and-Death Processes

The class of all continuous-time Markov chains has an important subclass formed by the birth-and-death processes. These processes are characterized by the property that whenever a transition occurs from one state to another, then this transition can be to a neighboring state only. Suppose that the state space is $S = \{0, 1, 2, \ldots, i, \ldots\}$, then transition, whenever it occurs from state i, can be only to a neighboring state $(i - 1)$ or $(i + 1)$.

A continuous-time Markov chain $\{X(t), t \in T\}$ with state space $S = \{0, 1, 2, \ldots\}$ and with rates

$$\begin{aligned} q_{i,i+1} &= \lambda_i \text{ (say)}, \quad i = 0, 1, \ldots, \\ q_{i,i-1} &= \mu_i \text{ (say)}, \quad i = 1, 2, \ldots, \\ q_{i,j} &= 0, \quad j \neq i \pm 1, \quad j \neq i, \quad i = 0, 1, \ldots, \quad \text{and} \\ q_i &= (\lambda_i + \mu_i), \quad i = 0, 1, \ldots, \quad \mu_0 = 0, \end{aligned}$$

is called

(i) a *pure birth process*, if $\mu_i = 0$ for $i = 1, 2, \ldots$,

(ii) a *pure death process*, if $\lambda_i = 0, i = 0, 1, \ldots$, and

(iii) a *birth-and-death-process* if some of the λ_i's and some of the μ_i''s are positive.

Using (1.3.12) we get the Chapman-Kolmogorov forward equations for the birth-and-death process.

For $i, j = 1, 2, \ldots$,

$$p_{ij}'(t) = -(\lambda_j + \mu_j) p_{ij}(t) + \lambda_{j-1} p_{i,j-1}(t) + \mu_{j+1} p_{i,j+1}(t) \quad (1.4.1)$$

and $$p_{i0}'(t) = -\lambda_0 p_{i0}(t) + \mu_1 p_{i,1}(t). \quad (1.4.2)$$

The boundary conditions are

$$p_{i,j}(0+) = \delta_{ij}, \quad i, j = 0, 1, \ldots. \quad (1.4.3)$$

Denote

$$P_j(t) = Pr\{X(t) = j\}, \quad j = 0, 1, \ldots, t > 0$$

and assume that at time $t = 0$, the system starts at state i, so that

$$P_j(0) = Pr\{X(0) = j\} = \delta_{ij}, \tag{1.4.4}$$

then

$$P_j(t) = p_{ij}(t),$$

and the forward equations can be written as

$$P_j'(t) = -(\lambda_j + \mu_j)P_j(t) + \lambda_{j-1}P_{j-1}(t) + \mu_{j+1}P_{j+1}(t), \; j = 1, 2, \ldots, \tag{1.4.5}$$

$$P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t). \tag{1.4.6}$$

Suppose that all the λ_i's and μ_i's are nonzero. Then the Markov chain is irreducible. It can be shown that such a chain is non-null persistent and that the limits

$$\lim_{t\to\infty} p_{ij}(t) = p_j$$

exist and are independent of the initial state i. Then Eqs. (1.4.5) and (1.4.6) become

$$0 = -(\lambda_j + \mu_j)p_j + \lambda_{j-1}p_{j-1} + \mu_{j+1}p_{j+1}, \quad j \geq 1 \tag{1.4.7}$$

$$0 = -\lambda_0 p_0 + \mu_1 p_1. \tag{1.4.8}$$

Define

$$\pi_j = \frac{\lambda_0\lambda_1\ldots\lambda_{j-1}}{\mu_1\mu_2\ldots\mu_j}, \quad j \geq 1, \quad \text{and}$$

$$\pi_0 = 1; \tag{1.4.9}$$

then the solution of the above can be obtained by induction. We have from (1.4.8)

$$p_1 = \left(\frac{\lambda_0}{\mu_1}\right) p_0 = \pi_1 p_0$$

and assuming $p_k = \pi_k p_0, k = 1, 2, \ldots, j$, we get from (1.4.7)

$$p_{j+1}\mu_{j+1} = \lambda_i \pi_j p_0, \quad \text{or}$$

$$p_{j+1} = \pi_{j+1} p_0.$$

Thus, if $\sum_{k=0}^{\infty} \pi_k < \infty$, then

$$p_j = \frac{\pi_j}{\sum \pi_k}, \quad j \geq 0. \tag{1.4.10}$$

Incidentally, $\sum \pi_k < \infty$ is a sufficient condition for the birth-and-death process to have all the states non-null persistent (and therefore for the process to be ergodic).

This process is of particular interest in queueing theory as several queueing systems can be modeled as birth-and-death processes. As an example, we consider the simple queue.

1.4.1 Special case: M/M/1 queue

For this queueing model

$$\lambda_j = \lambda, \quad i = 0, 1, 2, \ldots \quad \text{and}$$
$$\mu_i = \mu, \quad i = 1, 2, \ldots, \mu_0 = 0;$$

then $\pi_j = (\lambda/\mu)^j$ and $\sum \pi_k < \infty$ *iff* $(\lambda/\mu) < 1$, and then

$$\sum \pi_k = 1/[1 - (\lambda/\mu)]$$
$$p_j = [1 - (\lambda/\mu)](\lambda/\mu)^j, \quad j = 0, 1, 2, \ldots \tag{1.4.11}$$

1.4.2 Pure birth process: Yule-Furry process

If $\mu_i = 0, i \geq 0$, then we get a pure birth process; further, if $\lambda_i = i\lambda$, for all i, we get the Yule-Furry process for which

$$P_j'(t) = -j\lambda P_j(t) + (j-1)\lambda P_{j-1}(t), \quad j \geq 1, \quad \text{and}$$
$$P_0'(t) = 0 \tag{1.4.12}$$

1.5 Poisson Process

If $\mu_i = 0, i \geq 0, \lambda_i = \lambda$ for all i, then we get what is known as the homogeneous Poisson process with parameter λ. It is a pure birth process with constant rate λ.

The Poisson process can be used as a model of a large class of stochastic phenomena and is thus extremely useful from the point of view of application.

The Chapman-Kolmogorov forward equations are

$$P_j'(t) = -\lambda[P_j(t) - P_{j-1}(t)], \quad j \geq 1 \tag{1.5.1a}$$
$$P_0'(t) = -\lambda P_0(t). \tag{1.5.1b}$$

Let the boundary condition be

$$P_j(0) = \delta_{ij}.$$

The Eq. (1.5.1a) can be solved in a number of ways. Let us consider the method of generating function. Define

$$P(s,t) = \sum_{j=0}^{\infty} P_j(t)s^j, \tag{1.5.2}$$

(when the RHS converges); then

$$P(s,0) = s^i. \tag{1.5.3}$$

Assuming the validity of term-by-term differentiation, we get from (1.5.2)

$$\begin{aligned}\frac{\partial}{\partial t}P(s,t) &= \sum_{j=0}^{\infty}\frac{\partial}{\partial t}\{P_j(t)\}s^j \\ &= P_0'(t) + \sum_{j=1}^{\infty} P_j'(t)s^j.\end{aligned}$$

Multiplying (1.5.1a) by s^j and adding over $j = 1, 2, 3, \ldots$, we get

$$\frac{\partial}{\partial t}P(s,t) - P_0'(t) = -\lambda[P(s,t) - P_0(t) - sP(s,t)].$$

Using (1.5.1b), we have

$$\frac{\partial}{\partial t}P(s,t) = P(s,t)[\lambda(s-1)].$$

Solving, we get

$$\begin{aligned}P(s,t) &= Ce^{\lambda(s-1)t} \\ &= s^i e^{\lambda(s-1)t};\end{aligned} \tag{1.5.3a}$$

whence

$$\begin{aligned}P_j(t) &= \text{coeff. of } s^j \text{ in } \quad P(s,t) \\ &= e^{-\lambda t}\frac{(\lambda t)^{j-i}}{(j-i)!}, \quad j = i, i+1, \ldots, \\ &= 0, \quad j = 0, 1, \ldots, i-1.\end{aligned} \tag{1.5.4}$$

Since the Poisson process is a Markov chain $\{X(t), t \in (0,\infty)\}$ with stationary transition probabilities, we have

$$\begin{aligned}Pr\{X(t+s) - X(s) = k \mid X(s) = i\} &= Pr\{X(t+s) = i + k \mid X(s) = i\} \\ &= \frac{(\lambda t)^k}{k!}e^{-\lambda t}, \qquad i, k = 0, 1, \ldots; t, s \geq 0.\end{aligned} \tag{1.5.5}$$

We have defined the Poisson process as a birth process with constant birth rate. It can be introduced as a renewal process (as we shall see later in Section 1.7). A third way of defining the Poisson process is given below.

Let $N(t)$ denote the number of occurrences of a specified event in an interval of length t (i.e., during the time period, say, from 0 to t). Let

$$P_n(t) = Pr\{N(t) = n\}, \quad n = 0, 1, \ldots$$

We make the following postulates:

(1) *Independence.* The number of events occurring in two disjoint intervals of time are independent—that is, if $t_0 < t_1 < t_2, \ldots$, then the increments $N(t_1) - N(t_0), N(t_2) - N(t_1), \ldots$ are independent RVs.

(2) *Homogeneity in time.* The RV $\{N(t + s) - N(s)\}$ depends on the length of the interval $(t + s) - s = t$ and not on s or on the value of $N(s)$.

(3) *Regularity or orderliness.* In an interval of infinitesimal length h, the probability of *exactly one* occurrence is

$$P_1(h) = \lambda h + o(h)$$

and the probability of two or more occurrences is

$$\sum_{k=2}^{\infty} P_k(h) = o(h).$$

It follows that $P_0(h) = 1 - \lambda h + o(h)$. From the assumption of independence, we get

$$P_0(t + h) = P_0(t) P_0(h) = P_0(t)[1 - \lambda h + o(h)]$$

so that

$$\lim_{h \to o} \frac{P_0(t + h) - P_0(t)}{h} = -\lambda P_0(t) + \lim_{h \to 0} \frac{o(h)}{h}$$

or

$$P_0'(t) = -\lambda P_0(t).$$

We have

$$\begin{aligned} P_j(t + h) &= P_j(t) P_0(h) + P_{j-1}(t) P_1(h) + \sum_{r=2}^{\infty} P_{j-r}(t) P_r(h) \\ &= P_j(t)[1 - \lambda h + o(h)] + P_{j-1}(t)[\lambda h + o(h)] + o(h) \end{aligned}$$

so that

$$\lim_{h \to 0} \frac{P_j(t + h) - P_j(t)}{h} = -\lambda P_j(t) + \lambda P_{j-1}(t) + \lim_{h \to 0} \frac{o(h)}{h}$$

$$\text{or} \qquad P_j'(t) = -\lambda[P_j(t) - P_{j-1}(t)], \quad j \geq 1.$$

Thus, we get the same Chapman-Kolmogorov equations as given in (1.5.1a).

If $P_j(0) = \delta_{ij}$, then $P_j(t)$ is given by (1.5.4). When $P_j(0) = \delta_{0j}$ we get

$$P_j(t) = e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \quad j = 0, 1, 2, \ldots \tag{1.5.6}$$

Thus, $N(t)$ follows Poisson distribution with parameter λt—that is, $\{N(t), t \geq 0\}$ is a Poisson process with parameter λ (or rate λ). We have $E[N(t)] = \lambda t$ and var $[N(t)] = \lambda t$. We shall state below some important properties of the Poisson process. For proof see works on stochastic processes such as Karlin and Taylor (1975), Medhi (1994), and Ross (1980, 1983). For an account of the Poisson process, see Kingman (1993); also Serfozo (1990).

1.5.1 Properties of the Poisson process

(1) *Additive Property.* Sum of n independent Poisson processes with parameter $\lambda_i, i = 1, 2, \ldots, n$ is a Poisson process with parameter $\lambda_1 + \lambda_2 + \cdots + \lambda_n$.

(2) *Decomposition Property.* Suppose that $N(t)$ is the number of occurrences of a specified event and that $\{N(t), t \geq 0\}$ is a Poisson process with parameter λ. Suppose further that each occurrence of the event has a probability p of being recorded, and that recording of an occurrence is independent of other occurrences and also of $N(t)$. If $M(t)$ is the number of occurrences so recorded, then $\{M(t), t \geq 0\}$ is also a Poisson process with parameter λp; if $M_1(t)$ is the number of occurrences not recorded, then $\{M_1(t), t \geq 0\}$ is a Poisson process with parameter $\lambda(1-p)$. Further, $\{M(t), t \geq 0\}$ and $\{M_1(t), t \geq 0\}$ are independent.
The above implies that a *random selection* of a Poisson process yields a Poisson process.
In fact, a Poisson process can be decomposed into any number of independent Poisson processes—that is, a Poisson process is *infinitely divisible.*

(3) *Interarrival Times.* The interarrival times (i.e., the intervals) between two successive occurrences of a Poisson process with parameter λ are IID RVs that are exponential with mean $1/\lambda$.

(4) *Memoryless Property of Exponential Distribution.* Exponential distribution possesses what is known as a *memoryless* or *Markovian* property and is the only continuous distribution to possess this property. It may be stated as follows. Suppose that X has exponential distribution with mean $1/\lambda$; then

$$Pr\{X \geq x + y \mid X \geq x\} = Pr\{X \geq y\}$$

is independent of x, for the LHS equals

$$\begin{aligned} \frac{Pr\{X \geq x + y \text{ and } X \geq x\}}{Pr\{X \geq x\}} &= \frac{Pr\{X \geq x + y\}}{Pr\{X \geq x\}} \\ &= \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} \\ &= e^{-\lambda y} = Pr\{X \geq y\}. \end{aligned}$$

If the interval between two occurrences is exponentially distributed, then the memoryless property implies that the interval to the next occurrence is statistically independent of the time from the last occurrence and has exponential distribution with the same mean.

If τ is an arbitrary epoch in the interval (t_i, t_{i+1}) between the ith and $(i+1)$th occurrences of a Poisson process with parameter λ, then the distribution of the interval $(t_{i+1} - \tau)$ is independent of the elapsed time $(\tau - t_i)$ since the last occurrence and is exponential with mean $1/\lambda$.

In the queueing context, arrivals (or service completions) may be taken as occurrences; so in the case of the Poisson-exponential process, the above remarkable property leads to easily tractable and mathematically agreeable results.

(5) *Randomness Property.* Given that exactly one event of a Poisson process $\{N(t), t \geq 0\}$ has occurred by epoch T, then the time interval γ in $[0, T]$ in which the event occurred has uniform distribution in $[0, T]$. In other words,

$$Pr\{t < \gamma \leq t + dt \mid N(T) = 1\} = \frac{dt}{T}, \quad 0 < t < T.$$

This is also expressed by saying that an event of a Poisson process is a purely *random* event. The Poisson process is sometimes called a completely random process.

The preceding result holds in a more general case. This is stated below.

If an interval of length T contains exactly m occurrences of a Poisson process, then the joint distribution of the epochs at which these events occurred is that of m points uniformly distributed over an interval of length T. The result holds in case of the (more general) birth process of which the Poisson process forms a special class.

1.5.2 Generalization of the Poisson process

There are several directions in which the classical Poisson process can be generalized.

1.5.2.1 Poisson cluster process (compound Poisson process)

One of the postulates of the Poisson process is that at most one event can occur at a time. Now suppose that several events (i.e., a cluster of events) can occur simultaneously at an epoch of occurrence of a Poisson process $N(t)$ and that the number of events X_i in the ith cluster is a RV, X_is having independent and identical distributions

$$Pr\{X_i = j\} = p_j, \quad j = 1, 2, \ldots.$$

Then $M(t)$, the total number of events in an interval of length t, is given by

$$M(t) = \sum_{i=1}^{N(t)} X_i.$$

The stochastic process $\{M(t), t \geq 0\}$ is called a *compound Poisson process.* Its PGF is given by

$$G[P(s)] = \exp\{\lambda t[P(s) - 1]\},$$

where $P(s)$ is the PGF of X_i and $G(s)$ is the PGF of $N(t)$. We have

$$\begin{aligned} Pr\{M(t) = m\} &= \sum_{k=0}^{m} [Pr\{N(t) = k\} Pr\{X_i = m\}] \\ &= \sum_{k=0}^{m} e^{-\lambda t} \frac{(\lambda t)^k}{k!} p_m^{k*}, \end{aligned}$$

where p_m^{k*} is the probability associated with a k-fold convolution of X_i with itself.

We have

$$\begin{aligned} E\{M(t)\} &= \lambda t E\{X_i\} \quad \text{and} \\ \text{var}\{M(t)\} &= \lambda t E\{X_i^2\}. \end{aligned}$$

The compound Poisson process is useful in modeling queueing systems with batch arrival/batch service, exponential interarrival/service time, and independent and identical batch-sized distribution.

1.5.2.2 Nonhomogeneous Poisson process

The parameter λ in the classical Poisson process is assumed to be a constant, independent of time. Generalizations of the Poisson process arise when λ is assumed to be (i) a nonrandom function of time $\lambda(t)$ and (ii) a random variable.

Here it is assumed that the probability that arrival occurs between time t and time $t + \Delta t$, given that n arrivals occurred by time t, is equal to $\lambda(t)\Delta t + o(\Delta t)$, while the probability that more than one arrival occurs is $o(\Delta t)$. The resulting process is the so-called nonhomogeneous Poisson process $\{N(t), t \geq 0\}$. It can be shown that

$$\begin{aligned} p_n(t) &= Pr\{N(t) = n\} \\ &= \exp\left\{ - \int_0^t \lambda(x)dx \right\} \frac{\left[\int_0^t \lambda(x)dx\right]^n}{n!}, \quad n \geq 0. \end{aligned}$$

1.5.2.3 Random variation of parameter

Here we assume that λ is a random variable having PDF $f(\lambda), 0 \leq \lambda \leq \infty$. Thus,

$$p_n(t) = Pr\{N(t) = n\} = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} f(\lambda) d\lambda.$$

The case when the parameter λ of a Poisson process is a random function of time $\lambda(t)$ (and so is itself a stochastic process) leads to a doubly stochastic Poisson process. There are several situations where such generalizations of Poisson process may be realistic.

1.5.2.4 Truncated process

A simple generalization is truncation of the infinite domain of the Poisson process. This case arises in modeling a queueing system with waiting space limited to n; so arrivals that occur when the waiting space is full are not permitted and are lost to the system. This will be involved only in scaling the Poisson probabilities by a suitable scale factor.

1.5.3 Role of the Poisson process in probability models

The Poisson process and its associated exponential distribution possess many agreeable properties that lead to mathematically tractable results when used in probability models. Its importance is also due to the fact that occurrences of events in many real-life situations do obey the postulates of the Poisson process, and thus its use in probability modeling is considered realistic. An arrival process to a queueing system is often taken to be Poisson.

Consider an event and an interval of time during which the occurrences of the event happen. Suppose that the interval is subdivided into a large number of subintervals (say, n), and that p_i is the probability of occurrence of the event in the ith subinterval. Suppose further that the events occur independently of one another and that $\lambda = p_1 + \cdots + p_n$, while the largest of p_i tends to 0. Then the number of occurrences of the event in the interval tends in the limit to a Poisson distribution with mean λ. The Poisson distribution thus gives an adequate description of the cumulative effect of a large number of events, such that occurrence of an event in a small subinterval is improbable.

There are other contexts arising out of extreme value theory as well as information theory that provide justification of using the Poisson process in modeling.

Note: Rego and Szpankowski (1989) show that there is an equivalence between using entropy maximization with a two-moment constraint and assumption of exponential distribution in a certain queueing context.

1.6 Randomization: Derived Markov Chains

Let $\{X(t), t \geq 0\}$ be a continuous-time Markov chain with transition matrix Q and countable state space S. Assume that $X(t)$ is *uniformizable*—that is, the diagonal elements of Q are uniformly bounded. Let

$$\alpha = \sup_i q_i < \infty.$$

Then there exists a discrete-time Markov chain $\{Y_n, n = 0, 1, \ldots\}$ with state space S and TPM $P = (p_{ij})$ such that

$$\boldsymbol{P} = \frac{Q}{\lambda} + I, \tag{1.6.1}$$

where λ is any real number not less than α. Since $P(t) = e^{Qt}$(Eqn. 1.3.14), we have

$$\begin{aligned} P(t) &= e^{Qt} \\ &= e^{\lambda(P-I)t} \\ &= e^{-\lambda t} e^{\lambda P t} \\ &= e^{-\lambda t} \sum_{n=0}^{\infty} \frac{\lambda^n t^n}{n!} P^n, \end{aligned}$$

so that elementwise

$$p_{ij}(t) = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{\lambda^n t^n}{n!} p_{ij}^{(n)}, \quad t \geq 0, \quad i, j \in S. \tag{1.6.2}$$

We shall have from the above

$$\pi(t) = \pi(0) P(t) = \pi(0) e^{-\lambda t} \sum_{n=0}^{\infty} \frac{\lambda^n t^n}{n!} P^n \tag{1.6.3a}$$

or elementwise

$$\pi_j(t) = \pi(0) e^{-\lambda t} \sum_{n=0}^{\infty} \frac{\lambda^n t^n}{n!} p_{ij}^{(n)}. \tag{1.6.3b}$$

Another interesting fact is that there exists a Poisson process $\{N(t), t \geq 0\}$ with parameter λ such that Y_n and $N(t)$ are independent and that $\{X(t), t \geq 0\}$ and $\{Y_{N(t)}, t \geq 0\}$ are probabilistically identical—that is, we can write

$$X(t) \equiv Y_{N(t)}.$$

The converse also holds: if $X(t) = Y_{N(t)}$, then $Q = \lambda(P - I)$.

1.6.1 Markov chain on an underlying Poisson process (or subordinated to a Poisson process)

The above method of construction leads from a Markov process $\{X(t), t \geq 0\}$ to a derived Markov chain $\{Y_{N(t)}, t \geq 0\}$ by randomization of operational time through events of a Poisson process. For, we can obtain $p_{ij}(t)$ in terms of $p_{ij}^{(n)}$ by conditioning over the number of occurrences of the Poisson process $N(t)$ in $(0, t)$. Conditioning over the number of occurrences of the Poisson process with parameter λ over $[0, 1]$, we get

$$\begin{aligned} p_{ij}(t) &= Pr\{X(t) = j \mid X(0) = i\} \\ &= \sum_{n=0}^{\infty} Pr\{X(t) = j \mid X(0) = i, N(t) = n\} \\ &\quad \times Pr\{N(t) = n \mid X(0) = i\}. \end{aligned}$$

Now

$$Pr\{N(t) = n \mid X(0) = i\} = e^{-\lambda t}\left[\frac{(\lambda t)^n}{n!}\right]$$

and $Pr\{X(t) = j \mid X(0) = i, N(t) = n\}$ is the probability that the system goes from state i to state j in time t during which n Poisson occurrences took place. (That is, n transitions took place. Here, the time interval t is replaced by number of transitions). Thus,

$$Pr\{X(t) = j \mid X(0) = i, N(t) = n\} = p_{ij}^{(n)}.$$

Hence, we have

$$p_{ij}(t) = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} p_{ij}^{(n)}. \tag{1.6.4}$$

1.6.2 Equivalence of the two limiting forms

Let $\{Y_n, n \geq 0\}$ be an irreducible and aperiodic chain with finite state space S and TPM P. Then from the ergodic theorem (Theorem 1.1) we get that

$$\lim_{n \to \infty} p_{ij}^{(n)} = v_j, \quad i, j \in S$$

exists and is independent of i, and $V = \{v_1, v_2, \ldots\}$ is the invariant distribution given by

$$VP = V, \qquad Ve = 1. \tag{1.6.5}$$

The Markov process $\{X(t) = Y_{N(t)}, t \geq 0\}$ is also aperiodic and irreducible and has the same state space S. From the ergodic theorem (Theorem 1.4), we get that

$$\lim_{t\to\infty} p_{ij}(t) = u_j$$

exists and is independent of i. Further, $\boldsymbol{U} = \{u_1, u_2, \ldots\}$ is a probability vector and $\boldsymbol{U}$ is given as the solution of

$$\begin{aligned} &\boldsymbol{UQ} = \mathbf{0}, \quad \boldsymbol{Ue} = 1, \\ &\boldsymbol{UQ} = 0 \leftrightarrow \boldsymbol{U}[\lambda(P - I)] = \mathbf{0} \\ &\qquad \leftrightarrow \boldsymbol{UP} = \boldsymbol{U}. \end{aligned} \tag{1.6.6}$$

Thus, from (1.6.5) and (1.6.6), we get

$$\boldsymbol{U} \equiv \boldsymbol{V}.$$

In other words,

$$\lim_{t\to\infty} p_{ij}(t) = \lim_{n\to\infty} p_{ij}^{(n)}, \quad i, j \in S. \tag{1.6.7}$$

1.6.3 Numerical method

The numerical method is a subject in itself. We discuss the importance of the randomization technique in numerical analysis. This method of construction gives very useful formulas for computation of $p_{ij}(t)$ or $\pi_j(t)$—that is, transient probabilities of a uniformizable Markov process $\{X(t), t \geq 0\}$. Ross (1980) calls this method *uniformization*, though *randomization* appears to be a more generally used term. What is generally done in computational work is to choose a truncation point N and to set N to bound the error of truncation ε as follows. From (1.6.2),

$$p_{ij}(t) = \sum_{n=0}^{N} e^{-\lambda t}\frac{(\lambda t)^n}{n!} p_{ij}^{(n)} + \sum_{n=N+1}^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!} p_{ij}^{(n)}, \tag{1.6.8}$$

where N is so chosen that the second term is less than or equal to the desired control error ε. It would ensure that $p_{ij}(t)$ would be accurate to within ε. The same holds for the computation of $\pi_j(t) = Pr\{X(t) = j\}$.

Algorithms for computation have been developed by Grassman (1977) and Gross and Miller (1984a,b). These algorithms have been shown to be useful for computation of transient probabilities of many stochastic systems such as queueing, inventory, reliability, and maintenance systems. (Refer to Gross and Miller (1984a) for their SERT algorithm).

1.7 Renewal Processes

1.7.1 Introduction

We noted that the interarrival (or interoccurrence) times between successive events of a Poisson process are IID exponential random variables. A possible generalization is obtained by removing the restriction of exponential distribution and by considering that the interarrival times are IID random variables with an arbitrary distribution. The resulting process is called a renewal process.

Definition 1.3. Let X_n be the interval between the $(n-1)th$ and nth events of a counting process $\{N(t), t \geq 0\}$. Let $\{X_n, n = 1, 2, \ldots\}$ be a sequence of nonnegative IID random variables having distribution function F with mean μ. Then $\{N(t), t \geq 0\}$ is said to be a *renewal process* generated or induced by the distribution F. Assume that $N(t)$ is independent of X_i.

The discrete time process $\{X_n, n = 1, 2, \ldots\}$ also represents the same renewal process. Let

$$S_0 = 0, \qquad S_n = X_1 + \cdots + X_n, \quad n \geq 1.$$

Then

$$N(t) = \sup\{n : S_n \leq t\}. \tag{1.7.1}$$

If $S_n = t$ for some n, then a renewal is said to occur at time t. Thus, S_n gives the epoch of nth renewal. We have $F_n(x) = Pr\{S_n \leq x\}$, and $F_n = F^{n*}$ where F^{n*} is the n-fold convolution of F with itself. Assume that $E\{X_i\} = \mu$ exists and is finite. The function $M(t) = E\{N(t)\}$ is called the *renewal function* (which is a nonrandom function of t). When it exists, the derivative $M'(t) = m(t)$ is called the *renewal density* (*not* a PDF). The distribution of $N(t)$ is given by

$$\begin{aligned} p_n(t) &= Pr\{N(t) = n\} \\ &= Pr\{N(t) \geq n\} - Pr\{N(t) \geq (n+1)\} \\ &= Pr\{S_n \leq t\} - Pr\{S_{n+1} \leq t\} \\ &= F_n(t) - F_{n+1}(t). \end{aligned} \tag{1.7.2}$$

It can be easily verified that for X_n exponential, $\{N(t), t \geq 0\}$ is a Poisson process. The average number of renewals by time t equals

$$\begin{aligned} M(t) &= \sum_{n=0}^{\infty} n p_n(t) \\ &= \sum_{n=1}^{\infty} F_n(t) = \sum_{n=1}^{\infty} F^{n*}(t) \\ &= F(t) + \sum_{n=1}^{\infty} F^{(n+1)*}(t). \end{aligned} \tag{1.7.2a}$$

Now,

$$\sum_{n=1}^{\infty} F^{(n+1)*}(t) = \sum_{n=1}^{\infty} \int_0^t F^{n*}(t-x)dF(x)$$

$$= \int_0^t \left\{ \sum_{n=1}^{\infty} F^{n*}(t-x) \right\} dF(x)$$

assuming the validity of interchange of summation and integration operations. Thus,

$$M(t) = F(t) + \int_0^t M(t-x)dF(x). \tag{1.7.3}$$

The above is known as the *fundamental equation of renewal theory.*

Renewal theorems involving limiting behavior of $M(t)$ are interesting as well as important from the point of view of applications. (For details refer to any work on stochastic processes, such as Çinlar (1975), Karlin and Taylor (1975), Medhi (1994), and Ross (1983).)

1.7.2 Residual and excess lifetimes

We discuss below two RVs that arise in several situations. To a given $t > 0$, there corresponds a unique $N(t)$ such that

$$S_{N(t)} \leq t < S_{N(t)+1}, \tag{1.7.4}$$

that is, t lies in the interval $X_{N(t)+1}$ between $\{N(t)\}$th and $\{N(t)+1\}$th renewals.

The RV $Y(t) = S_{N(t)+1} - t$ (which is the interval between t and the renewal epoch after t) is called the *residual lifetime* or *forward-recurrence time* at t.

The RV $Z(t) = t - S_{N(t)}$ (which is the interval between t and the last renewal epoch before t) is called the *spent lifetime* or *excess lifetime* or *backward-recurrence time* at t.

Note that

$$Y(t) + Z(t) = S_{N(t)+1} - S_{N(t)} = X_{N(t)+1} \tag{1.7.5}$$

is the total life.

These RVs arise in various queueing contexts. The RV $Z(t)$ denotes the elapsed time between t and the last arrival before t or between t and the commencement of the last service before t depending on whether X_i denotes the interarrival or service time. Similarly, $Y(t)$ can be interpreted. We consider now the distribution of $Y(t)$ and $Z(t)$. We have

$$Pr\{Y(t) \leq x\} = F(t+x) - \int_0^t [1 - F(t+x-y)]dM(y), \quad x > 0$$

$$= 0, \quad x \leq 0. \tag{1.7.6}$$

If F is not a lattice distribution, then the limiting distribution Y of $Y(t)$ is given by

$$Pr\{Y \leq x\} = \lim_{t\to\infty} Pr\{Y(t) \leq x\} = \frac{1}{\mu}\int_0^x [1 - F(y)]dy, \quad x \geq 0. \quad (1.7.7)$$

Again,

$$Pr\{Z(t) \leq x\} = \begin{cases} 0, & x \leq 0, \\ F(t) - \int_0^{t-x} [1 - F(t-y)]dM(y), & 0 < x \leq t, \\ 1, & x > t, \end{cases} \quad (1.7.8)$$

and if F is not a lattice distribution, then the limiting distribution Z of $Z(t)$ is given by

$$\begin{aligned} Pr\{Z \leq x\} &= \lim_{t\to\infty} Pr\{Z(t) \leq x\} \\ &= \frac{1}{\mu}\int_0^x [1 - F(y)]dy, \quad x \geq 0 \\ &= 0, x < 0. \end{aligned} \quad (1.7.9)$$

Assumption of finite mean μ ensures that the above is a proper distribution (in both the cases). Z is also called stationary excess distribution.

When these exist, the two limiting distributions Y and Z are identical. It can be easily verified that for exponential X_i, the distributions of $Y(t)$ and $Z(t)$ are again exponential with the same mean $\mu = E(X_i)$.

Suppose that $m_r = E(X_i^r)$ exist for $r = 1, 2$. Then

$$E\{Y\} = E\{Z\} = \frac{m_2}{2\mu} = \frac{m_2}{2m_1}. \quad (1.7.10)$$

If F is a lattice distribution, then the distributions of $Y(t)$ and $Z(t)$ have no limits for $t \to \infty$ except in some special cases.

Note: Higher moments of Y (or Z) can be found in terms of those of X (see Problem 1.26).

1.8 Regenerative Processes

Let $\{X(t), t \geq 0\}$ be a stochastic process with countable state space $S = \{0, 1, 2, \ldots\}$. Suppose that there exists an epoch t_1 such that the continuation of the process beyond t_1 is a probabilistic replica of the whole process starting at $0(=t_0)$. Then this implies the existence of epochs $t_2, t_3, \ldots (t_i > t_{i-1})$

having the same property. Such a process is known as a regenerative process. If $T_n = t_n - t_{n-1}, n = 1, 2, \ldots$, then $\{T_n, n = 1, 2, \ldots\}$ is a renewal process.

A renewal process is regenerative, with T_i representing the time of the ith renewal.

Another example of a regenerative process is provided by what is known as an *alternating renewal process.* Such a process can be envisaged by considering that a system can be in one of two possible states—say, 0 and 1—that is, having $S = \{0, 1\}$. Initially, it is at state 0 and remains at that state for a time Y_1, and then a change of state to state 1 occurs in which it remains for a time Z_1, after which it again goes to state 0 for a time Y_2 and then goes to state 1 for a time Z_2 and so on. That is, its movement could be denoted by $0 \to 1 \to 0 \to 1 \ldots$. The initial state could be 1, in which case the movement could be denoted by $1 \to 0 \to 1 \to 0 \ldots$.

Suppose that $\{Y_n\}, \{Z_n\}$ are two sequences of IID random variables and that Y_n and Z_n need not be independent. Let

$$T_n = Y_n + Z_n, \quad n = 1, 2, \ldots.$$

Then at time T_1 the process restarts itself, and so also at times $T_2, T_3 \ldots$. The interval T_n denotes a complete cycle, and the process restarts itself after each complete cycle. Let

$$E\{Y_n\} = E\{Y\}, \qquad E\{Z_n\} = E\{Z\}.$$

Then the long-run proportions of time that the system is at states 0 and 1 are given, respectively, by

$$p_0 = \lim_{t \to \infty} Pr\{X(t) = 0\} = \frac{E\{Y\}}{E\{Y\} + E\{Z\}} \tag{1.8.1}$$

and

$$p_1 = \lim_{t \to \infty} Pr\{X(t) = 1\} = \frac{E\{Z\}}{E\{Y\} + E\{Z\}}$$

$$= 1 - p_0. \tag{1.8.2}$$

1.8.1 Application in queueing theory

The results (1.8.1) and (1.8.2) have an important application in queueing theory. Consider a single-server queueing system such that an arriving customer is immediately taken for service if the server is free, but joins a waiting line if the server is busy. The system can be considered to be in two states (idle or busy) according to whether the server is idle or busy. The idle and busy states alternate and together constitute a cycle of an alternating renewal process. A busy period starts as soon as a customer arrives before an idle server and ends at the instant when the server becomes free for the first time.

The epochs of commencement of busy periods are regeneration points. Let I_n and B_n denote the lengths of nth idle and busy periods, respectively, and let

$$E\{I_n\} = E\{I\} \quad \text{and}$$
$$E\{B_n\} = E\{B\}. \tag{1.8.3}$$

Then the long-run proportion of time that the server is idle equals

$$p_0 = \frac{E\{I\}}{E\{I\} + E\{B\}}, \tag{1.8.4}$$

and the long-run proportion of time that the server is busy equals

$$p_1 = \frac{E\{B\}}{E\{I\} + E\{B\}}. \tag{1.8.5}$$

In particular, if the arrival process is Poisson with mean λt, then it follows (from its lack of memory property) that an idle period is exponentially distributed with mean $1/\lambda$—that is, $E(I) = 1/\lambda$. Then when p_0 or p_1 is known, $E(B)$ can be found.

The case of the alternating renewal process can be generalized to cover cyclical movement of more than two states. Suppose that the state space of the process $\{X(t), t \geq 0\}$ is $S = \{0, 1, \ldots, m\}$ and its movement from initial state 0 is cyclic as $0 \to 1 \to 2 \cdots m \to 0 \ldots$, and that τ_k is the duration of sojourn at state k, having mean $\mu_k = E\{\tau_k\}, k = 0, 1, \ldots, m$. Then

$$p_k = \lim_{t\to\infty} Pr\{X(t) = k\} = \frac{\mu_k}{\sum_{i=0}^{m} \mu_i}, \quad k = 0, 1, \ldots, m. \tag{1.8.6}$$

1.9 Markov Renewal Processes and Semi-Markov Processes

We shall now consider a kind of generalization of a Markov process as well as a renewal process. Let $\{X(t), t \geq 0\}$ be a Markov process with discrete countable state $S = \{0, 1, 2, \ldots\}$, and let $t_0 = 0, t_1, t_2 \cdots (t_i < t_{i+1})$ be the epochs at which transitions occur. The sequence $\{X_n = X(t_n + 0), n \geq 0\}$ forms a Markov chain, and the transition intervals $T_n = t_n - t_{n-1}, n = 1, 2, \ldots$ are distributed as independent exponential variables having means that may depend on the state of X_n.

We generalize the situation as follows: Suppose that the transitions $\{X_n, n \geq 0\}$ of the process $\{X(t), t \geq 0\}$ constitute a Markov chain but the transition intervals $T_n, n = 0, 1, \ldots$ have an independent arbitrary distribution and that the mean may depend not only on the state of X_n but also on the state of X_{n+1}.

The process $\{X(t), t \geq 0\}$ is then no longer Markovian. The two-dimensional process $\{X_n, t_n n \geq 0\}$ is called a Markov renewal process with state space S. Here

$$\begin{aligned} Pr\{X_{n+1} = j, T_{n+1} \leq t \mid X_0 = x_0, \ldots, X_n = i, T_0, T_1, \ldots, T_n\} \\ = Pr\{X_{n+1} = j, T_{n+1} \leq t \mid X_n = i\} \\ = Q_{ij}(t), \quad \text{say } i, j \in S. \end{aligned} \tag{1.9.1}$$

Let

$$p_{ij} = \lim_{t \to \infty} Q_{ij}(t) \quad \text{and} \quad F_{ij}(t) = \frac{Q_{ij}(t)}{p_{ij}} \quad \text{and} \tag{1.9.2}$$

$$Y(t) = X_n \text{ on } t_n \leq t < t_{n+1}. \tag{1.9.3}$$

Then $\{Y(t), t \geq 0\}$ is called a *semi-Markov process*, and the Markov chain $\{X_n, n \geq 0\}$ is called the *embedded Markov* chain of $\{X(t), t \geq 0\}$. $Y(t)$ gives the state of the process at its most recent transition. The chain $\{X_n, n \geq 0\}$ has TMP (p_{ij}). $F_{ij}(t) = Pr\{T_{ij} \leq t\}$ is the distribution function of T_{ij}, the conditional transition time (or sojourn time) at state i given that the next transition is to state j. If τ_k is the unconditional waiting time at state k, then $\tau_k = \sum_j p_{ij} T_{ij}$.

For example, a pure birth process is a special type of Markov renewal process having

$$\begin{aligned} Q_{ij}(t) &= 1 - e^{a_i t}, \quad j = i + 1, \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Then

$$\begin{aligned} p_{ij} &= 1, \quad j = i + 1, \\ &= 0 \quad \text{otherwise, and} \\ F_{ij}(t) &= Q_{ij}(t), \qquad T_i = T_{ij}, \quad j = i + 1. \end{aligned}$$

A Markov renewal process becomes a Markov process when the transition times are independent exponential and are independent of the next state visited. It becomes a Markov chain when the transition times are all identically equal to 1. It reduces to a renewal process if there is only one state and then only transition times becomes relevant. Semi-Markov processes are used in the study of certain queueing systems. Let $p_k = \lim_{t \to \infty} Pr\{Y(t) = k\}$ be the long-run proportion of time the semi-Markov process is at state k. Suppose that the embedded Markov chain $\{X_n, n = 0, 1, 2, \}$ is irreducible, aperiodic, and, if denumerable, recurrent nonnull. Then the limiting probabilities

$$v_j = \lim_{n \to \infty} p_{ij}^{(n)}$$

exist and are given as the unique non-negative solution of

$$v_j = \sum_{k \in s} v_k p_{kj}, \quad j \in S.$$

Then we shall have

$$p_k = \frac{v_k \mu_k}{\sum_{j \in S} v_j \mu_j}, \tag{1.9.4}$$

where $\mu_k = E\{\tau_k\}$ is the expected sojourn time in state k. One can get this result by extending the result (1.8.6) through an intuitive argument. For a formal proof, see Medhi (1994).

Problems

1.1. The transition probability matrix of a Markov chain with three states 0, 1, 2, is given by

$$\begin{pmatrix} 0.4 & 0.5 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

and the initial distribution is $(0.6, 0.3, 0.1)$. Find (i) $Pr(X_2 = 3)$, and (ii) $Pr\{X_3 = 1, X_2 = 0, X_1 = 2, X_0 = 0\}$. Find the invariant measure of the chain.

1.2. A chain with $S = \{1, 2, 3, \}$ has TPM

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_1 & p_2 & p_3 \end{pmatrix}, \quad p_i > 0, \ \sum p_i = 1.$$

Examine the nature of the states. Find P^n.

1.3. Find the invariant measure of a chain with $S = \{0, 1, 2, \ldots, m-1\}$ and a doubly stochastic transition probability matrix.

1.4. Consider a Markov chain with $S = \{1, 2, 3, 4\}$ and TPM

$$\begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Is the chain irreducible? Verify that states 1 and 2 are recurrent. Find μ_1 and μ_2. $(\mu_1 = 5/3, \mu_2 = 5/2)$.

1.5. Show that for a Markov chain with a finite state space S, the probability of staying forever among the transient states is zero.

1.6. Show that if state j is transient, then

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty \quad \text{for all} \quad i \in S.$$

1.7. Show that a transient state cannot be reached from a persistent state.

1.8. Consider a service facility having a limited waiting space for m customers, including the one being serviced. The server serves one customer, if any, at epochs $0, 1, 2, \ldots$. Assume that the number of arrivals in the intervals $(k, k+1)$ is given by an IID random variable A with $Pr(A = n) = p_n, \sum p_n = 1$. Assume further that arrivals that occur when the waiting space is full leave the system and do not return. Denote by X_n the number of customers present at time n, including the one being served, if any. Show that $\{X_n, n \geq 0\}$ is a Markov chain and find its TPM.

1.9. In what is called a *gambler's ruin problem*, consider a gambler who with capital a agrees to play a series of games with an adversary having a capital $b (a + b = c$, the total capital). The probability of the gambler winning one game (and with it, one unit of money) is p, and that of losing one unit is $q = 1 - p$. (There is no draw.) Suppose that successive games are independent. If X_n is the gambler's fortune at time n (at time of the nth game), show that $\{X_n, n = 0, 1, 2, \ldots\}$ is a Markov chain. Write down its TPM. Is the chain irreducible? Examine the nature of the states of the chain.

1.10. Consider the Markov chain with $S = \{0, 1, \ldots, m\}$, such that

$$\begin{aligned} p_{0,0} &= q, \quad p_{01} = p, \\ p_{i,i-1} &= q, \quad p_{i,i+1} = p, \quad i = 1, 2, \ldots, m-1 \\ p_{m,m-1} &= q, \quad p_{m,m} = p, \quad 0 < p, q < 1, \end{aligned}$$

where $p + q = 1$. (Each of the transitions between other pairs of states has probability 0.) Show that the chain is irreducible and aperiodic. Find the limiting distribution $\boldsymbol{V}$.

1.11. Consider the Markov chain of Example 1.3.

(a) Use (1.2.21) to find $\sum_j j v_j$.

(b) For $p_n = q^n p, n = 0, 1, 2, \ldots, p + q = 1$, examine the nonnull persistence of the chain. Find $V(s)$ in this case. Consider the particular case $q = \rho/(1 + \rho), \rho < 1$.

1.12. Suppose that customers arrive at a bank in accordance with a Poisson process at the rate of two per minute. Find the probability that the number of customers that arrive during a 10-minute period is (i) exactly 20, (ii) greater than 20, and (iii) between 10 and 20.

1.13. Suppose that customers arrive at a certain service facility center in accordance with a Poisson process $\{N(t), t \geq 0\}$ having parameter λ. A customer can make a preliminary inquiry as to whether he or she actually needs the facility. Suppose that the proportion of customers actually needing the service facility is $p(0 < p < 1)$. If $M(t)$ gives the number of customers actually needing service, show that $\{M(t), t \geq 0\}$ is again a Poisson process with parameter λp.

1.14. Suppose that customers arrive at a service counter in accordance with a Poisson process at a rate of three per minute. Find the probability that the interval between two successive arrivals is (i) between one and three minutes and (ii) less than one-third of a minute.

1.15. Suppose that messages arrive at a telephone switch board, the interarrival time of messages being exponential with mean 10 minutes. Find the probability that the number of messages received during the five afternoon hours (1–6 P.M.) is (i) exactly 24, (ii) more than 24, and (iii) nil.

1.16. If $X_i, i = 1, \ldots, n$ are IID exponential RVs with parameter a, then show that $S_n = X_1 + \cdots + X_n$ has gamma distribution with PDF

$$f_{a,n}(x) = \frac{a^n x^{n-1} e^{-ax}}{\Gamma(x)}, \quad x > 0$$
$$= 0, \quad x \leq 0.$$

Find $E\{S_n\}$ and var $\{S_n\}$.

1.17. Suppose that X and Y are independent exponential RVs with parameters a and b, respectively. Show that

(a) $W = X + Y$ has PDF

$$f(x) = \frac{ab(e^{-ax} - e^{-bx})}{b - a}, \quad x > 0, \quad a \neq b$$
$$= 0, \quad x \leq 0.$$

(b) $Z = \min(X, Y)$ is exponential with parameter $a + b$.

(c) $M = \max(X, Y)$ has PDF

$$f(x) = ae^{-ax} + be^{-bx} - (a + b)e^{-(a+b)x}, \quad x > 0$$
$$= 0, \quad x \leq 0.$$

(d)

$$Pr\{X \leq Y\} = \frac{a}{(a+b)}.$$

1.18. If X is an exponential RV, then show that

$$E\{X \mid X > y\} = y + E\{X\} \quad \text{for all} \quad y > 0;$$

that is, $E\{X - y \mid X > y\} = E(X)$, independent of y.

1.19. A piece of equipment is subject to random shocks that occur in accordance with a Poisson process with rate λ. The equipment fails due to the cumulative effect of k shocks. Show that the duration of the lifetime T of the equipment has gamma distribution with PDF $f_{\lambda,k}(x)$. Note that T is the interval between k occurrences of a Poisson process.

1.20. Consider that two independent series of events A and B occur in accordance with Poisson processes with parameters a and b, respectively. Show that the number N of occurrences of the event A between two *successive* occurrences of the event B has geometric distribution with mass function

$$Pr\{N = n\} = (1-q)q^n, \qquad q = \frac{a}{(a+b)}, \quad n = 0, 1, 2, \ldots$$

1.21. Consider a Poisson process with parameter λ. Given that n events happen by time t, show that the PDF of the time of occurrence T_k of the kth event $(k < n)$ is given by

$$f(x) = \frac{n!}{(k-1)!(n-k)!} \frac{x^{k-1}}{t^k} \left(1 - \frac{x}{t}\right)^{n-k}, \quad 0 < x < t,$$
$$= 0, \quad x \geq t.$$

Show that $E(T_k) = kt/(n+1)$.

1.22. Suppose that a queueing system has m service channels. The demand for service arises in accordance with a Poisson process with rate a, and the service time distribution has exponential distribution with parameter b. Suppose that the service system has no storage facility—that is, a demand that arises when all m channels are busy is rejected and is lost to the system. Let $X(t)$ be the number of busy service channels (number of demands) at time t. Show that $\{X(t), t \geq 0\}$ is a continuous-time Markov chain. Determine the infinitesimal generator. (See problem 1.8 for the same queue with discrete service time.)

1.23. **The three-state process.** Suppose that an automatic machine can be in three states: working (state 0), failed in mode 1 (state 1), or failed in mode 2 (state 2). Suppose that a failed machine in either mode cannot go to another failed mode. (That is, transitions from state 1 to 2 and from state 2 to 1 are not possible.) Suppose that $X(t)$ denotes the state (condition)

of the machine at time t and that the failed times are IID exponential with rates $a_i, i = 1, 2$, and the repair times are IID exponential with rates $b_i, i = 1, 2$. Show that $\{X(t), t \geq 0\}$ is a continuous-time Markov chain. Find the Q-matrix. (State any assumption that you make.)

1.24. Assume that the lifetime X of a device is random having DF $F(x)$. Show that the expected remaining life of a device aged y (which has already attained age y) is given by

$$E\{X - y \mid X > y\} = \frac{\int_y^\infty [1 - F(t)]dt}{1 - F(y)}.$$

In particular, for the exponential lifetime, this is equal to $E(X)$, independent of y. (See problem 1.18.)

1.25. Let $\{N(t), t \geq 0\}$ be a renewal process induced by a RV X. Show that for large t

$$E\{N(t)\} \simeq \frac{t}{E(X)} \quad \text{and}$$

$$\text{var}\{N(t)\} \simeq \frac{\text{var}(X)}{[E(X)]^3} t.$$

1.26. Let Y be the stationary forward-recurrence time of a random variable X. Suppose that $m_n = E(X^n)$ exists for $n = 1, 2, \ldots$. Then show that

$$E\{Y^n\} = \frac{m_{n+1}}{(n+1)m_1}.$$

1.27. Prove that the limiting joint distribution of the residual lifetime $Y(t)$ and spent lifetime $Z(t)$ of a random variable with DF $F(x)$ and finite mean μ is given by

$$Pr\{Y(t) > y, Z(t) > z\} = \frac{1}{\mu} \int_{y+z}^\infty [1 - F(u)]du, \quad y > 0, \quad z > 0.$$

1.28. **Renewal-reward process.** Consider a renewal process $\{X_n, n = 1, 2, \ldots\}$. Suppose that renewal epochs are $t_0 = 0, t_1, \ldots$, and that $N(t)$ is the number of renewals by time t; associate a RV $Y_i (i = 1, 2, \ldots)$ with renewal epoch $t_i (i = 1, 2, \ldots)$. (A reward or cost is associated with each renewal, the amount being given by a RV Y_i for ith renewal.)
Let

$$Y(t) = \sum_{i=1}^{N(t)} Y_i.$$

Then the stochastic process $\{Y(t), t \geq 0\}$ is called a renewal-reward process. Suppose that $E(X_n) = E(X)$ and $E(Y_n) = E(Y)$ are finite.

Then show that (a) with probability 1,

$$\lim_{t\to\infty} \frac{Y(t)}{t} \to \frac{E(Y)}{E(X)};$$

and that (b)

$$\lim_{t\to\infty} \frac{E\{Y(t)\}}{t} \to \frac{E(Y)}{E(X)}.$$

The relation (b) gives the long-run average reward (cost) per unit time in terms of $E(X)$ and $E(Y)$.

References and Further Reading

Bhat, U. N. (1984). *Elements of Applied Stochastic Processes*, 2nd ed., Wiley, New York.

Çinlar, E. (1975). *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ.

Grassman, W. K. (1977). Transient solutions on Markovian queueing systems. *Comp. Opns. Res.* **4**, 47–53.

Gross, D., and Miller, D. R. (1984a). The randomization technique as a modeling tool and solution procedures for transient Markov processes. *Opns. Res.* **32**, 343–361.

Gross, D., and Miller, D. R. (1984b). Multiechelon repairable-item provisioning in a time-varying environment using a randomization technique. *Naval Res. Log. Qrly.* **31**, 347–361.

Karlin, S., and Taylor, H. M. (1975). *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York.

Kaufman, L. (1983). Matrix methods for queueing problems. *SIAM J. Sci. Stat. Comp.*, **4**, 525–552.

Kingman, J. F. C. (1993). *Poisson Process*, Oxford University Press, Oxford.

Lal, R., and Bhat, U. N. (1987). Reduced systems in Markov chains and their applications in queueing theory. *Queueing Systems* **2**, 147–172; Correction **4**, 93.

Medhi, J. (1994). *Stochastic Processes*, 2nd ed. J. Wiley & Sons, New York & Wiley Eastern Ltd. (now New Age International Publishers), New Delhi; 1st ed (1982).

Prabhu, N. U. (1965). *Stochastic Processes*, Macmillan, New York.

Rego, V., and Szpankowski, W. (1989). The presence of exponentiality in entropy maximized $M/GI/1$ queues. *Comp. Opns. Res.* **16**, 441–449.

Ross, S. M. (1980). *Introduction to Probability Models*, 2nd ed., Academic Press, New York.

Ross, S. M. (1983). *Stochastic Processes*, Wiley, New York.

Seneta, E. (1981). *Non-negative Matrices and Markov Chains*, 2nd ed., Springer, New York.

Serfozo, R. F. (1990). Point Processes in *Handbooks in Operations Research and Management Science*, Vol. 2, pp. 1–94 (Eds. D. P. Heyman and M. J. Sobel), North-Holland, Amsterdam.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ.

Tang, Chi-Nang, and Yeung, R. W. (1999). A Graph theoretic approach to Queueing Analysis, Part I & II. *Comm. Statist.—Stochastic Models.* **15**, 791–870.

Whitt, W. (1982). Approximating a point process by a renewal process. Two basic methods. *Opns. Res.* **30**, 125–147.

Whitt, W. (1983). Untold horrors of the waiting time: what the equilibrium distribution will never tell about the queue length process. *Mgmt. Sci.* **29**, 395–408.

Whitt, W. (2002). *Stochastic-Process Limits (An Introduction to Stochastic-Process Limits and Their Application to Queueing)*, Springer-Verlag, New York.

Queueing Systems: General Concepts

2.1 Introduction

The origin of queueing theory dates back to 1909, when Agner Krarup Erlang (1878–1929) published his fundamental paper on congestion in telephone traffic [for a brief account, see Saaty (1957), and for details on his life and work, see Brockmeyer *et al.* (1948)].

In addition to formulating in analytic form several practical problems arising in telephony and solving them, Erlang laid solid foundations for queueing theory in terms of the nature of assumptions and techniques of analysis; these are being routinely used to this day even in wider areas of modern communications and computer systems. In a way, Erlang was a pioneer in the applications of analytical methods to operational problems. His studies appear to mark the beginning of the study of Operations Research.

Kendall (1951, 1953) was the pioneer who viewed and developed queueing theory from the perspective of stochastic processes.

The literature on queueing theory and the diverse areas of its applications has grown tremendously (exponentially, as claimed by some writers) over the years. For a bibliography of books and survey papers, refer to Prabhu (1987), Takagi (1991), Dshalalow (1995, 1997), and Takagi and Boguslavsky (1990). The last named includes titles of fundamental monographs first written in Russian. Hlynka gives a list of books in his Web site.

Queueing theory is the mathematical study of "queues" or "waiting lines." A queue is formed whenever the demand for service exceeds the capacity to provide service at that point in time. A queueing system is composed of customers or units needing some kind of service who arrive at a service facility where

such service is provided, join a queue if service is not immediately available, and eventually leave after receiving service. There are also cases where customers leave the system without joining the queue or leave without receiving service even after waiting for some time.

The terms *customer* and *server* are generic ones. Customers are those who need some kind of service and arrive at a facility where such service is available. A mechanism that performs the kind of service on customers or units fed into it is called a server or a service channel—for example, customers at a bank or reservation counter, calls arriving at a telephone keyboard, machines with a repairman needing repair, merchandise for shipment at a yard, and so on. Jobs arriving at a component in a computer center are also regarded as customers, and the component of the computing system (such as CPU, drum, disk, line printer, etc.) where such a facility is provided is considered the server. A customer receiving service is said to be in service. If upon arrival a "customer" finds the server busy, it forms or joins a queue.

2.1.1 Basic characteristics

The basic characteristics of a queueing system are as follows:

(i) The Input or arrival pattern of customers;

(ii) The pattern of service;

(iii) The number of servers or service channels;

(iv) The capacity of the system; and

(v) The queue discipline.

We describe these characteristics below.

2.1.2 The input or arrival pattern of customers

The *input pattern* means the manner in which the arrivals occur. It is specified by the interarrival time between any two consecutive arrivals. A measure usually considered is the average length of the interarrival time or its reciprocal, the average number of arrivals per some unit of time. The input pattern also indicates whether the arrivals occur singly or in groups or batches. If in batches, the manner in which these batches are constituted is also to be covered in the input pattern. The interarrival time may be deterministic, so it is the same between any two consecutive arrivals, or it may be stochastic, when its distribution is also to be specified. Sometimes an arrival may not join the queue being discouraged by the length of the queue or being debarred from joining the system because the waiting space, when limited, is filled to maximum capacity. Again, the arrivals may occur from an infinite source or sometimes from a finite source, with the same units circulating in the system—

that is, machines (in an establishment) coming for repair whenever they fail. There may also be several classes of customers with different arrival rates.

2.1.3 The pattern of service

By the *pattern of service*, we mean the manner in which the service is rendered. It is specified by the time taken to complete a service. The time may be constant (*deterministic*) or it may be stochastic. If it is stochastic, the pattern specification involves the distribution of service time of a unit. A measure typically considered is provided by the average time required to serve a unit or by the average number of units served per some unit of time. Sometimes service may be rendered in bulks or batches, as in the case of an elevator, instead of personalized service of one at a time. In this case, the manner of formation of batches for service also has to be specified.

2.1.4 The number of servers

A system may have a single server or a number of parallel servers. An arrival who finds more than one free server may choose at random any one of them for receiving service. If he finds all the servers busy, he joins a queue common to all the servers. The first customer from the common queue goes to the server who becomes free first. This kind of situation is common—for example, in a bank or at a ticket counter.

There may also be situations where there is a separate queue in front of each service facility, as in the case of a supermarket. There also arise cases of ordered entry when an arrival has to try to find a free server in the order the servers are arranged.

Unless otherwise stated, we will use the term *multiple server system* (with a number of parallel channels) to indicate a system having a common queue, with the head of the queue going to the first free server.

2.1.5 The capacity of the system

A system may have an infinite capacity—that is, the queue in front of the server(s) may grow to any length. Against this there may be limitation of space, so that when the space is filled to capacity, an arrival will not be able to join the system and will be lost to the system. The system is called a *delay system* or a *loss system*, according to whether the capacity is infinite or finite. If finite, it will have to be specified by the number of places available for the queue as well as for the one(s) being served, if any.

2.1.6 The queue discipline

The queue discipline indicates the manner in which the units are taken for service. The usual queue discipline is first come, first served, or FCFS (first in first out, FIFO), though sometimes there are other service disciplines, such

as last come, first served (which happens sometimes in case of messages) or service in random order.

When arrivals occur in batches and service is offered individually, then the manner in which customers arriving in a batch are arranged for service is also to be indicated.

There are also such disciplines as processor-sharing, usually adopted in computer systems with a number of terminals.

Sometimes customers may be of several kinds with a definite order of priority for receiving service with preemptive (or nonpreemptive) service discipline.

The demand for service is often expressed as $a = \lambda/\mu$, the mean number of arrivals per unit time taken as the mean service time. This quantity is called *offered load* (or *traffic intensity*); it is a dimensionless quantity and is expressed in "erlang" (in honor of A. K. Erlang, the father of queueing theory). This offered load is a measure of what the customers want.

The quantity $\rho = a/c$ when the numbers of servers is $c(\geq 1)$ is the offered load per server and is called *carried load.* It is also called *utilization factor* or server utilization.

For a single-server system, $a = \rho$ gives the long-run proportion (fraction) of time the server is busy. These have to be suitably defined in case of batch arrivals and/or bulk service and also in finite systems with limited waiting space or limited input source.

Unless otherwise stated, by a system we shall mean a single-server system with arrivals from an infinite source, with unlimited waiting space and with FIFO queue discipline. The interarrival (as well as service times) will be assumed to be mutually independent and also to be independent of one another.

2.2 Queueing Processes

The analysis of a queueing system with fixed (deterministic) interarrival and service times does not present much difficulty. We shall be concerned with models or systems where one or both (interarrival and service times) are stochastic. Their analyses will involve a stochastic description of the system and related performance measures, as discussed below.

(1) Distribution of the number $N(t)$ in the system at time t (the number in the queue and the one being served, if any). $N(t)$ is also called the queue length of the system at time t. By the *number* in the system (queue), we will always mean the *number of customers* in the system (queue).

(2) Distribution of the waiting time in the queue (in the system), the time that an arrival has to wait in the queue (remain in the system). If W_n denotes the waiting time of the nth arrival, then of interest is the distribution of W_n.

(3) Distribution of the virtual waiting time $W(t)$—the length of time an arrival has to wait had he arrived at time t.

(4) Distribution of the busy period being the length (or duration) of time during which the server remains busy. The busy period is the interval from the moment of arrival of a unit at an empty system to the moment that the channel becomes free for the first time. The busy period is a random variable.

From a complete description of the above distributions, various performance measures of interest are obtained.

The problems studied in queueing theory may be grouped as:

(i) Stochastic behavior of various random variables, or stochastic processes that arise, and evaluation of the related performance measures;

(ii) Method of solution—exact, transform, algorithmic, asymptotic, numerical, approximations, etc.;

(iii) Nature of solution—time dependent, limiting form, etc.;

(iv) Control and design of queues—comparison of behavior and performances under various situations, as well as queue disciplines, service rules, strategies, etc.; and

(v) Optimization of specific objective functions involving performance measures, associated cost functions, etc.

Analysts and operations researchers generally will be involved with these types of problems. But in order to study such problems, one will have to study first the types of problems enumerated under (i)–(iii).

2.3 Notation

The notation introduced by Kendall (1951) is generally adopted to denote a queueing model. It consists of the specifications of three basic characteristics: the input, the service time, and the number of (parallel) servers. Symbols used to denote some of the common formulations are as follows:

M	Exponential interarrival (Poisson input) and service time distribution (having Markov property)
E_k	Erlang-k distribution
H	Hyperexponential distribution
PH	Phase-type distribution
D	Deterministic (constant)(interarrival or service time)
G	Arbitrary (general) distribution

Thus, the notation $M/G/1$ denotes a queue or model with Poisson input, general service time distribution, and a single server. Two more descriptors are added when needed: the fourth one to denote the capacity of the system and the fifth one to denote the size of the (finite) source from which the arrivals occur. Thus, the $M/G/1$ model stands for an $M/G/1/\infty$ model. $G/G/c/K/N$ refers to the c-server model with (arbitrary) general interarrival and service time distributions, the space before the servers (in the system) being limited to K (including the ones being served, if any), and the N being the size of the finite source from which the arrivals occur. The description is suitably modified to cover more complicated models or models with other characteristics.

2.4 Transient and Steady-State Behavior

Denote by $N(t)$ the number in the system (the number in the queue plus the number being served, if any) at time t measured from a fixed initial moment ($t = 0$) and its probability distribution by

$$p_n(t) = Pr\{N(t) = n\}, \quad n = 0, 1, 2, \ldots. \text{ Then}$$
$$p_i(0) = 1, \quad (p_j(0) = 0, j \neq i)$$

implies that the number of customers at the initial moment was i (where i could be $0, 1, 2, \ldots$.). For a complete description of the stochastic behavior of the queue-length processes $\{N(t), t \geq 0\}$, we need to find a time-dependent solution $p_n(t), n \geq 0$. It is often difficult to obtain such solutions. Or even when found, these may be too complicated to handle. For many practical situations, however, one needs the equilibrium behavior—that is, the behavior when the system reaches an equilibrium state after being in operation for a sufficiently long time. In other words, one is often interested in the limiting behavior of $p_n(t)$ as $t \to \infty$. Denote

$$p_n = \lim_{t \to \infty} p_n(t), \quad n = 0, 1, 2, \ldots$$

whenever the limit exists. Thus, p_n is the limiting probability that there are n in the system, irrespective of the number at time 0. Whenever the limit exists, the system is said to reach a steady (or equilibrium) state, and p_n is called the steady-state probability that there are n in the system. It is independent of time and $\{p_n\}$ is said to have a steady-state or stationary or equilibrium distribution.

It usually turns out that p_n is equal to the long-run proportion of time that the system contains exactly n customers. In particular, p_0 denotes the proportion of time that the system is empty. It follows that

$$\sum_{n=0}^{\infty} p_n = 1;$$

this is called the normalizing condition.

It is necessary to know the condition(s) under which the limit exists. This will be discussed at appropriate places.

We consider two other sets of *limiting* probabilities $\{a_n, n \geq 0\}$ and $\{d_n, n \geq 0\}$ defined as follows:

a_n = probability that arrivals (arriving customers) find n in the system when they arrive;

d_n = probability that departures (departing customers) leave n in the system when they depart.

It turns out that a_n is the long-run proportion of customers who, when they arrive, find n in the system, and d_n is the long-run proportion of customers who, when they depart, leave n in the system. The three quantities p_n, a_n, and d_n need not always be all equal.

We prove the following theorem.

Theorem 2.1. Burke's Theorem

In any queueing system in which arrivals and departures occur one by one and that has reached equilibrium state,

$$a_n = d_n \quad \textit{for all } n \geq 0.$$

Proof: Consider that an arrival will see, on arrival, n in the system; then the number in the system will increase by 1 and will go from n to $n + 1$. Again, a departure will leave n in the system, implying that the number in the system will decrease by 1 and will go from $n + 1$ to n. In any interval of time T, the number of transitions A from n to $n + 1$ and the number of transitions B from $n + 1$ to n will differ at most by 1; in other words, either $A = B$ or $A \sim B = 1$. Then for large T, the rates of transitions A/T and B/T will be equal. Thus, on the average, arrivals and departures always see the same number of customers, which means that $a_n = d_n$ always and for every $n \geq 0$. ■

Note: It may be noted that a_n (and d_n) is in general different from p_n; that is, the long-run proportion of time that arrivals find n in the system is not, in general, equal to the long-run proportion of time that there are n customers in the system. This is stated by saying that arrivals, in general, do not see time averages. In one important case, however, $a_n = p_n$; that is, when the arrivals are from a Poisson process. This is discussed in Section 2.7.1.

2.5 Limitations of the Steady-State Distribution

In certain situations, the conditions for the existence of stationary distribution exist and a stationary or equilibrium distribution of the behavior of the queueing process can be obtained. However, even when this happens, there can be, in general, several stochastic processes having the same stationary distribution.

Whitt (1983) points out the danger of using only the stationary distribution to describe the stochastic behavior of a queue with detailed study of a $GI/M/1$ queue in that context.

Consider that the stationary distribution is known or given. The problem of finding all queue-length processes or all arrival processes associated with a given stationary distribution is an *inverse* problem. Such inverse problems arise in several other situations (Karr and Pittenger, 1978; Keller, 1976). In the study of Markov chains, the inverse problem is the one of characterizing the class of Markov transition matrices with a given stationary distribution of the chain (Karr, 1978).

In order to be specific about the behavior of queueing processes, Whitt (1983) points to the necessity of examining the transient behavior of some characteristics associated with the queueing process in addition to the stationary distribution, which by itself is not enough. The transient behavior that can be considered may be of various first-passage times such as the busy period. That is, in addition to stationary distribution, such other distributions may be considered. The transient behavior is also to be studied because of its importance in the cost-benefit analysis of an operating system. An example considered by Whitt is the allocation of buffer by a central processor where the stationary distribution of buffer content may be used for determination of the number of buffers required, but the stochastic fluctuations will indicate the load on the central processor for buffer allocation. Cohen (1982) considers relaxation time to describe and measure the rate at which a stochastic process approaches steady state.

2.6 Some General Relationships in Queueing Theory

There are certain useful statements, relationships in queueing theory that hold under fairly general conditions. Though rigorous mathematical proofs of such relations are somewhat complicated, intuitive and heuristic proofs are simple enough and have been known for a long time (Morse, 1958). It has been argued (Krawkowski, 1973; Stidham 1974) that *conservation methods* could very well be applied to supply simple proofs of some of these relations. Conservation principles have played a fundamental role in physical and engineering sciences as well as in economics and so on. Similar principles may perhaps be applied in obtaining relations for queueing systems in steady state. Some such relations that hold for systems in *steady state* are given below. The most important one is

$$L = \lambda W \tag{2.6.1}$$

where λ is the mean arrival rate, L is the expected number of units in the system, and W is the expected waiting time in the system in steady state. Denote the

expected number in the queue and the expected waiting time in the queue in steady state by L_Q and W_Q, respectively. These are related by a similar formula:

$$L_Q = \lambda W_Q. \tag{2.6.2}$$

The fundamental insight of the truth of the above is due to Morse (1958). His student, Little (1961), gave a rigorous proof of the relation, and so the relation is known as Little's formula. Jewell (1967) gave a proof based on renewal theory; Eilon (1969) gave the following simple proof. For a historical account, see Whitt (1991). It is stated that Jewell's (1967) proof was the first complete proof; it does not require steady-state conditions.

Eilon's proof of $L = \lambda W$. This proof does not depend (i) on the arrival or service-time distributions, (ii) on the number of servers in the system, or (iii) on the queue discipline.

Consider Fig. 2.1. The top line gives the cumulative number of arrivals and the bottom line the cumulative number of departures from the system. The vertical distances between the two lines gives the number of customers present in the system at that instant, while the horizontal distance denotes the waiting time in the system (waiting time in the queue plus service time).

Suppose that the system has been in operation for some time and that it has settled down to a steady-state condition. Consider a time interval T that may include none or more than one busy period. Let

$A(T)$ = total number of arrivals during the period T

$B(T)$ = area between the two horizontal systems of lines

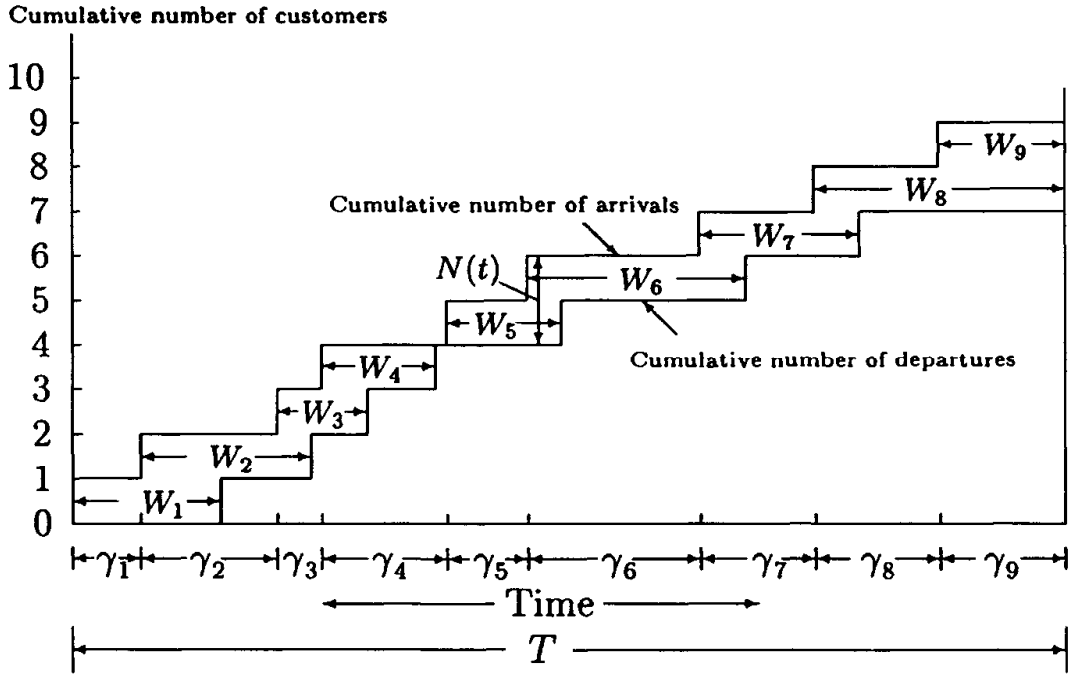


Figure 2.1 Cumulative number of arrivals and departures of a queueing system.

$$
\begin{aligned}
&= \text{total waiting time (in the system) of all the customers who arrive during } T \\
&= W_1 + W_2 + \cdots \\
\lambda(T) &= \text{mean arrival rate during } T \\
&= \frac{A(T)}{T} \\
W(T) &= \text{mean waiting time in the system during } T \\
&= \frac{B(T)}{A(T)} \\
L(T) &= \text{mean number of customers in system during } T \\
&= \frac{B(T)}{T}
\end{aligned}
$$

We have

$$L(T) = \frac{B(T)}{T} = \frac{B(T)}{A(T)} \frac{A(T)}{T} = W(T)\lambda(T).$$

Suppose that limits exist as $T \to \infty$ and are given by

$$\lim_{T\to\infty} \lambda(T) = \lambda \quad \text{and}$$
$$\lim_{T\to\infty} W(T) = W.$$

Then a limit for $L(T)$ as $T \to \infty$ also exists and is given by

$$L = \lim_{T\to\infty} L(T)$$

and the three limits satisfy the relation

$$L = \lambda W,$$

which is Little's formula for corresponding L and W in the system.

To get $L = \lambda W$ for L and W in the queue, we refer to the bottom line as representing departures from the *queue* and not from the system; then $L(T)$ and $W(T)$ refer to those in the queue. Proceeding as before, we get

$$L_Q = \lambda W_Q.$$

What we have shown in Fig. 2.1 is that the queue discipline is FCFS. A little reflection would reveal that the total waiting time $B(T)$ of all the customers who arrive during the interval T will remain unaltered under different service disciplines (and so $B(T)$ will not be affected by the service discipline). Thus the result holds irrespective of the service discipline, and also of the number of servers.

This completes the proof. ■

This result, of great generality, holds irrespective of the form of interarrival and service time distributions and discrepancy that may be within the system. It holds under some very general conditions for any system, provided it is in steady-state. Ramalhoto *et al.* (1983) give a survey of contributions on this subject, together with applications to various situations (for a review, see Whitt (1991)). The principle of customer conservation is that the frequency of entries into a service channel equals the frequency of departures from the channel. The frequency of arrivals is λ, and the frequency of departures is $\mu(1 - p_0)$. Thus, for a $GI/G/1$ queue, we have

$$\lambda = (1 - p_0)\mu \quad \text{or} \quad p_0 = 1 - \rho. \tag{2.6.3}$$

The average number of busy servers in a $GI/G/c$ queue is $c\rho = \lambda/\mu$.

There are other relations that hold under somewhat restrictive conditions. For example, for a single-server system with Poisson input, the Pollaczek-Khinchin formula holds. Some of these results are derived later.

From (2.6.1) and (2.6.2) it follows that for a $GI/G/c$ queue

$$L - L_Q = \lambda(W - W_Q); \tag{2.6.4}$$

we know that $W - W_Q$ is the expected service time, which equals $1/\mu$; and $L - L_Q$ is the expected number in service or the expected number of busy servers. Thus, the expected number of busy servers equals $c\rho$, and the expected number of idle servers equals $c - c\rho = c(1 - \rho)$.

Remarks:

(1) The formula $L = \lambda W$ is valid in great generality. It relates customer average waiting time W to time-average queue length L, given an arrival rate λ. It applies to other stochastic models besides queues.

(2) The result $L = \lambda W$ is very general. What we have considered in the simple proof is for customers. The result also holds for *servers*; in this case, L is the average number of busy servers, and W is the average service time. $c\rho$ equals λ/μ in a c-server system (with $\rho = \lambda/c\mu$).

(3) Heyman and Stidham (1980) show that similar relationships exist between more general customer and time averages, which are represented by the formula $H = \lambda G$. For motivation, consider an example provided by Glynn and Whitt (1989) relating to migration of salmon upstream. As the stream narrows, a queue of salmon (salmon ladder) is created. The amount of food consumed by each fish in the queue can be modeled as a stochastic process. Then H may be taken as the average amount of food consumed per fish in the queue among the first n fish (throughout all time) and G as the average amount of food consumed per time in the queue by time t (by all fish).

(4) The relation $L = \lambda W$ is a special case of the relation $H = \lambda G$, which embodies the full relationship between time averages and customer averages.

An extension of $H = \lambda G$ to represent lump costs as well as cost rates is given by Glynn and Whitt (1989).

(5) A general relationship of the type $H = \lambda G$ is also known as *Rate Conservation Law* (RCL) (Miyazawa, 1990); for a recent survey on RCL, see Miyazawa (1994).

(6) Attempts to generalize Little's formula has been mainly in two directions: (i) to weaken the sufficient conditions for the validity of the formula and (ii) to relate higher-order moments of the number in the system and sojourn time.

(7) There are other extensions of $L = \lambda W$ besides $H = \lambda G$. The extensions cover continuous, distributional, ordinal, and central limit theorem versions (see Glynn and Whitt (1986a,b)). For continuous analogue, refer to Rolski and Stidham (1983).

(8) For an extension to distributional form, see Keilson and Servi (1988, 1990) and Bertsimas and Nakazato (1995).

A form of *Distributional Law* is as follows:

Let N be the number in the system and W be the response time (queueing plus service times), then

$$\begin{aligned} & P_N(s) = W^*(\lambda - \lambda s) \\ \text{where} \quad & P_N(s) = E(s^N) \text{ is the pgf of } N \\ \text{and} \quad & W^*(s) = E[e^{-sW}] \text{ is the LST of } W. \end{aligned}$$

Differentiating the above relation w.r.t. s and then putting $s = 1$, one gets the Little's Law.

The distributional form holds for a wide variety of settings. Some of these are:

(a) When the reference is number N_Q in queue and waiting time W_Q in queue for the $GI/G/s$ model; also for the number N in the system and response time W in case of the models $GI/G/1$ and $GI/D/s$ (and not for $GI/G/s$).

(b) For $GI/G/1$ vacation models with exhaustive service as well as priority queues with only one priority class.

(9) For an $M/G/c$ queue

$$\begin{aligned} & E\{L^r\} = \lambda^r E\{W^r\}/r! \quad r = 1, 2, \ldots \\ \text{and} \quad & E\{L_{(r)}\} = \lambda^r E\{W^r\}, \\ \text{where} \quad & L_{(r)} = L(L-1)\ldots(L-r+1), \quad r = 1, 2, \ldots \end{aligned}$$

(See Brumelle, 1972.)

2.7 Poisson Arrival Process and Its Characteristics

In Chapter 1, we discussed the role of the Poisson process in the modeling of stochastic phenomena. Arrival processes in several situations can be taken as Poisson processes. For example, telephone calls received at a switchboard, arrival of customers in a bank, arrival of jobs at a CPU in a computer center, arrival of telecom messages, breakdown of machines in a machine shop with a large number of machines, and many other types of arrivals to a service center usually can be modeled as Poisson processes—that is, arrivals can be considered as events that occur in accordance with a Poisson process.

We mention here yet another interesting type of study. A service facility in a queue may receive input from a number of different sources. It would be reasonable, therefore, to postulate that the input or arrival process to the service facility is the superposition of a number of component processes that are nearly independent. Albin (1982), Whitt (1982), and Newell (1984) have studied the queue-length behavior of systems of the type $\sum_{i=1}^{n} GI_i/G/1$ where the input process is the superposition of n independent renewal processes. Albin has done simulation studies of queueing systems of the type $\sum GI_i/M/1$ with exponentially distributed service time and with an input process that is the superposition of n independent renewal processes each with rate λ/n. She observes that as n increases, the average queue length for such a system approaches that of an $M/M/1$ system. However, for fixed n the difference between the corresponding characteristic of the two systems increases as ρ increases from 0.5 to 0.9.

Newell (1984) studied in detail some of the qualitative properties of the $\sum GI_i/G/1$ system. He observes that as $n \to \infty$, under certain conditions, superposition of renewal processes behaves like a Poisson process. He further shows that for the average queue length the approach to the limiting $M/G/1$ behavior requires that $n(1-\rho)^2 \gg 1$.

These studies further strengthen the basis of assumption of the Poisson process as an arrival process.

2.7.1 PASTA: Poisson arrivals see time averages

Poisson arrivals have an interesting property that such arrivals behave like *random* arrivals. That is, an observer from a Poisson arrival stream sees or finds the same system state distribution as a *random* observer having nothing to do with the system (say from outside).

For Poisson arrivals systems

$$a_n = p_n \quad \text{for all } n \geq 0$$

provided the state of the system changes by at most one (that is, there is no bulk arrival nor bulk service). The result holds also in the transient state—that is, $a_n(t) = p_n(t)$, for $n \geq 0, t \geq 0$.

Let $A(t, t+\delta)$ be the number of arrivals in the infinitesimal interval $(t, t+\delta)$. We have

$$\begin{aligned} a_n(t) &= \lim_{\delta \to 0} Pr\{N(t) = n \mid \text{an arrival occurred just after instant } t\} \\ &= \lim_{\delta \to 0} Pr\{N(t) = n \mid A(t, t+\delta) = 1\} \\ &= \lim_{\delta \to 0} \frac{Pr\{N(t) = n, A(t, t+\delta) = 1\}}{Pr\{A(t, t+\delta) = 1\}} \\ &= \lim_{\delta \to 0} \frac{Pr\{A(t, t+\delta) = 1 \mid N(t) = n\} \cdot Pr\{N(t) = n\}}{Pr\{A(t, t+\delta) = 1\}} \end{aligned}$$

By the postulate of *independence* of Poisson process (see Section 1.5),

$$Pr\{A(t, t+\delta) = 1 \mid N(t) = n\} = Pr\{A(t, t+\delta) = 1\}$$

so that

$$a_n(t) = \lim_{\delta \to 0} P\{N(t) = n\} = p_n(t).$$

It follows that $a_n = p_n$ in steady state, for all $n \geq 0$.

The requirement is that future arrivals are independent of the current number in the system, which is satisfied by the Poisson arrival process.

This property can also be stated as **"Poisson Arrivals See Time Averages"** (**PASTA**). In a queue with Poisson arrivals, the limiting proportion of arrivals that find the system in some state n is equal to the limiting proportion of *time* the system spends in that state n. By **PASTA** is meant the equality of these two limiting fractions.

A simple proof of **PASTA** is due to Wolff (1982) under the lack of anticipation assumption (LAA).

Let $N(t)$ be the number in the system in a queue where the arrival process $A(t)$ is Poisson. Let B be an arbitrary set in the state space of $N(t)$. Denote

$$\begin{aligned} U(t) &= 1, \quad \text{if } N(t) \in B \\ &= 0, \quad \text{otherwise} \\ V(t) &= \left(\frac{1}{t}\right) \int_0^t U(s)\, ds, \\ Y(t) &= \int_0^t U(s)\, dA(s), \quad \text{and} \\ Z(t) &= \frac{Y(t)}{A(t)}. \end{aligned}$$

Then $V(t)$ is the fraction of time during $[0, t]$ that $N(t)$ is in state B and $Z(t)$ is the fraction of arrivals in $[0, t]$ that find $N(t)$ in state B. Then Wolff shows

that, under the lack of anticipation assumption, $t \to \infty$.

$$V(t) \to V(\infty) \text{ with probability } (w.p.)1$$

$$\text{iff} \quad Z(t) \to V(\infty) \quad \text{w.p. } 1.$$

Note: An elementary example of PASTA is given by Wolff (1989). It refers to a Poisson arrival queue with a single server with no waiting space (except the one under service, if any). It is shown that the fraction of time the server is busy (as seen by an outside observer) equals the fraction of arrivals who find the server busy.

An example, where $a_n \neq p_n$, relating to non-Poisson arrival is given here.

Non-Poisson Arrival system. Consider a single-server queueing situation where the arrival process is not Poisson. An interesting example is given in Bertsekas and Gallager (1994). Consider that the interarrival times are uniformly distributed between 2 minutes and 4 minutes and the service time is of constant duration 1 minute. Then an arriving customer will always find the system empty—that is, $a_0 = 1, a_n = 0$ $(n \geq 1)$. Now what will an outside arrival find? Here, $\mu = 1$ and

$$\frac{1}{\lambda} = \frac{1}{2} \int_2^4 x \, dx = 3,$$

so that utilization factor $\rho = Pr\{\text{server is busy}\} = \frac{1}{3}$.

An outside observer will find either 1 customer in the system with probability ρ or will find zero customer in the system with probability $(1 - \rho)$, so that $p_0 = \frac{2}{3}$ and $p_1 = \frac{1}{3}, p_n = 0, n \geq 2$. Thus, for this non-Poisson arrival, $a_n \neq p_n$.

The average number of customers in the system (as found by the outside observer) equals

$$1 \cdot \rho + 0 \cdot (1 - \rho) = \rho = \frac{1}{3}.$$

Notes:

(1) The result (PASTA) may be generalized to the case of a nonstationary Poisson arrival process with rate $\lambda(t)$.

(2) It is the independent-increments property of the Poisson process that really accounts for PASTA.

(3) Instances do occur when non-Poisson arrivals also see time averages (Burke, 1976).

(4) PASTA does not hold in case of quasi-random input. When the arrivals occur from a *finite* source, though the interarrival times are exponential RV,

$a_n \neq p_n$. An example is considered in Chapter 3 in connection with a finite system.

(5) It holds whether the system is a delay system or a loss system—for example, $M/G/1/K$.

(6) Niu (1984) studies the interesting question of when we have inequalities between the two limiting proportions and considers three types of interarrival-time distributions in this connection. This problem has been studied earlier by König *et al.*, Marshall and Wolff, Mori, Rolski, Stoyan, Whitt, and many others.

(7) Discrete time analogue of PASTA has also been considered.

2.7.2 ASTA: arrivals see time averages

It is also known that some non-Poisson arrivals see time averages. Thus Poisson arrival is sufficient but not necessary for

$$a_n = p_n$$

to hold (under LAA). Arrivals that see time averages (or ASTA): This topic has been discussed for a long time by several researchers, including Descloux (1967), Franken *et al.* (1981), Cooper (1990), Wolff (1982, 1989), Melamed and Whitt (1990), and König *et al.* (1983). A sufficient condition for ASTA to hold under Weak Lack of Anticipation Assumption (WLAA) (which is weaker than LAA) has been given in Melamed and Whitt (1990). They also give a NASC for ASTA in a stationary framework under what is known as Lack of Bias Assumption (LBA).

An elementary example of (queueing) in which ASTA holds for non-Poisson arrivals is given in König *et al.* (1983). It is of a $GI/M/1$ queue, in which the interarrival time distribution $A(x)$ has some specific property.

Anti-PASTA

There is another related question of interest: When does ASTA imply that the associated counting process is Poisson? It is called anti-PASTA (refer to Green and Melamed (1990), Melamed and Whitt (1990), and Melamed and Yao (1995)).

References and Further Reading

Albin, S. L. (1982). On Poisson approximations for superposition of arrival processes in queues. *Mgmt. Sci.* **28**, 127–137.

Bacceli, F., and Bremaud, P. C. (1994). *Elements of Queueing Theory*, Springer, New York.

Bertsekas, D., and Gallager, R. (1994). *Data Networks*, 2nd ed. (1st ed., 1987), Prentice-Hall, Englewood Cliff, NJ.

Bertsimas, D., and Nakazato, D. (1995). The general distributional Little's law and its applications. *Opns. Res.* **43**, 298–310.

Bocharov, P. P., and Pechinkin, A. V. (1995). *Theory of Queues*, Peoples' Friendship Univ. of Russia Publ., Moscow.

Brockmeyer, E., Halstorm, H. L., and Johnson, A. (1948). *The Life and Works of A. K. Erlang* (translation of the Danish Academy of Sciences, no. 2). The Copenhagen Telephone Company, Copenhagen. Second Edition (1960). *Acta Polytechnics Scandivanica (Applied Maths. & Computing Machinery Series AP* 287).

Brumelle, S. L. (1972). A generalization of $L = \lambda W$ to moments of queue lengths and waiting times. *Opns. Res.* **20**, 1127–1136.

Burke, P. J. (1976). Proof of a conjecture on the interarrival time distribution in an $M/M/1$ queue with feedback. *IEEE Trans. Com.* **24**, 575–576.

Cohen, J. W. (1982). *The Single Server Queue*, 2nd ed., North-Holland, Amsterdam.

Cooper, R. B. (1990). Queueing Theory in *Handbooks of Operations Research and Management Science* (Eds. D. P. Heyman and M. J. Sobel), vol. 2, pp. 469–518, North Holland, Amsterdam (contains a list of 214 References that includes books published in 1981–1989).

Descloux, A. (1967). On the validity of a particular subscriber's view. *Proc. Fifth Intl. Conf. Teletraffic Congress*, New York.

Dshalalow, J. H. (1995). An anthology of classical queueing methods in *Advances in Queueing* (Ed. J. H. Dshalalow), CRC Press, Boca Raton, pp. 1–42 (contains a list of 181 books on Queueing Theory, 74 books on other topics related to Queueing Theory, 234 books on Performance Evaluation & Computer Systems etc., and 116 Survey Papers).

Dshalalow, J. H. (1997). Queueing systems with state dependent parameters in *Frontiers in Queueing* (Ed. J. H. Dshalalow), CRC Press, Boca Raton, pp. 61–116 (contains a bibliography of 277 books).

Eilon, S. (1969). A simple proof of $L = \lambda W$. *Opns. Res.* **17**, 915–916.

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt. Tiddskrift Matematik* B **20**, 33–39.

Franken, P., König, D., Arndt, U., and Schmidt, V. (1981). *Queues and Point Processes*, Akademic-Verlag, Berlin.

Glynn, P. W., and Whitt, W. (1986a). A central-limit-theorem version of $L = \lambda W$. *Queueing Systems* **1**, 191–215.

Glynn, P. W., and Whitt, W. (1986b). Sufficient conditions for functional limit-theorem versions of $L = \lambda W$. *Queueing Systems* **1**, 279–287.

Glynn, P. W., and Whitt, W. (1989). Extensions of the queueing relation $L = \lambda W$ and $H = \lambda G$. *Opns. Res.* **37**, 634–644.

Green, L., and Melamed, B. (1990). An anti PASTA result for Markovian systems. *Opns. Res.* **38**, 173–175.

Heyman, D. P., and Stidham, S., Jr. (1980). The relation between customer average and time average in queues. *Opns. Res.* **28**, 983–994.

Hlynka, M. (2002). A list of 220 Books on Queueing Theory (Web site).

Jewell, W. S. (1967). A simple proof of $L = \lambda W$. *Opns. Res.* **17**, 1109–1116.

Karr, A. F. (1978). Markov chains and processes with a prescribed invariant measure. *Stoch. Proc. Appl.* **7**, 277–290.

Karr, A. F., and Pittenger, A. O. (1978). The inverse balayage problem for Markov chains. *Stoch. Proc. Appl.* **7**, 165–178.

Keilson, J., and Servi, L. D. (1988). A distributional form of Little's law, *Opns. Res. Letters* **7**, 223–227.

Keilson, J., and Servi, L. D. (1990). The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Opns. Res. Letters* **9**, 239–247.

Keller, J. B. (1976). The inverse problem. *Am. Math. Monthly* **83**, 107–118.

Kendall, D. G. (1951). Some problems in the theory of queues. *J.R.S.S.B* **13**, 151–185.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Ann. Math. Statist.* **24**, 338–354.

Kendall, D. G. (1964). Some recent work and further problems in the theory of queues. *Theo. Prob. Appl.* **9**, 1–15.

Kingman, J. F. C. (1993). *Poisson Process*, Oxford University Press, Oxford.

Konig, D., Miyazawa, M., and Schmidst, V. (1983). On the identification of Poisson arrivals in queues with coinciding time stationary and customer stationary state distributions. *J. Appl. Prob.* **20**, 860–871.

Krawkowski, M. (1973). Conservation methods in queueing theory. *Rev. Fran. d'Auto Inf. Rech. Oper.* 7e Annee, **V-1**, 3–20.

Little, J. D. C. (1961). A proof of the queueing formula $L = \lambda W$. *Opns. Res.* **9**, 383–387.

Louchard, G., and Latouche, G. (Eds.) (1983). *Probability Theory and Computer Science. Part II Stochastic Modeling: Queueing Models*, Academic Press, New York.

Melamed, B., and Whitt, W. (1990). On arrivals that see time averages. *Opns. Res.* **38**, 156–172.

Melamed, B., and Yao, D. D. (1995). The ASTA property in *Advances in Queueing* (Ed. J. H. Dshalalow), CRC Press, Boca Raton, FL, pp. 195–224.

Miyazawa, M. (1990). Derivation of Little's law and related formulas by rate conservation law with multiplicity. Department of Inf. Sc., Science, University of Tokyo.

Miyazawa, M. (1994). Rate conservation law—a Survey. *Queueing Systems* **5**, 1–58.

Morse, P. M. (1958). *Queues, Inventories and Maintenance*, Wiley, New York.

Newell, G. F. (1971). *Applications of Queueing Theory*, 2nd ed., 1982, Chapman & Hall, London.

Newell, G. F. (1984). Approximations for superposition arrival processes in queues. *Mgmt. Sci.* **30**, 623–632.

Niu, S.-C. (1984). Inequalities between arrival averages and time averages in stochastic processes arising from queueing theory. *Opns. Res.* **32**, 785–795.

Prabhu, N. U. (1986). Editorial introduction. *Queueing Systems* **1**, 1–4.

Prabhu, N. U. (1987). A bibliography of books and survey papers on queueing systems: theory and applications. *Queueing Systems* **2**, 393–398.

Prabhu, N. U. (1997). *Foundations of Queueing Theory*, Kluwer Academic Publications, New York.

Rolski, T., and Stidham, S., Jr. (1983). Continuous versions of the queueing formulas $L = \lambda W$ and $H = \lambda G$. *Opns. Res. Letters* **2**, 211–215.

Ramalhoto, M. F., Amaral, J. A., and Cochito, M. T. (1983). A survey of J. Little's formula. *Inter. Stat. Review* **51**, 255–278.

Saaty, T. L. (1957). A. K. Erlang. *Opns. Res.* **5**, 293–294.

Stidham, S., Jr. (1972). $L = \lambda W$: a discounted analogue and a new proof. *Opns. Res.* **20**, 1115–1126.

Stidham, S., Jr. (1974). A last word on $L = \lambda W$. *Opns. Res.* **22**, 417–421.

Strauch, R. E. (1970). When a queue looks the same to an arriving customer as to an observer. *Mgmt. Sci.* **17**, 140–141.

Takagi, H. (1991). *Queueing Analysis, A Foundation of Performance Evaluation, Vol. 1. Vacation and Priority Systems* Part I, North Holland, Amsterdam (contains also a list of 230 books that contain material on queueing theory).

Takagi, H., and Boguslavsky, L. B. (1991). A supplementary bibliography on queueing analysis and performance evaluation. *Queueing Systems* **8**, 313–322.

Whitt, W. (1982). Approximating a point process by renewal process I: two basic methods. *Opns. Res.* **30**, 125–147.

Whitt, W. (1983). Untold horrors of the waiting room: what the equilibrium distribution will never tell about the queuelength process. *Mgmt. Sci.* **29**, 395–408.

Whitt, W. (1991). A review of $L = \lambda W$ and extensions, *Queueing Systems* **9**, 235–268 (contains a list of 73 references).

Wolff, R. W. (1982). Poisson arrivals see time averages. *Opns. Res.* **30**, 223–231.

Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*, Prentice Hall, Englewood Cliffs, NJ.

Birth-and-Death Queueing Systems: Exponential Models

3.1 Introduction

Many simple but interesting queueing systems can be studied through birth-death processes, which have been discussed in Section 1.4. In such a process transitions take place from one state only to a neighboring state. With an arrival there is a transition from the state $k(\geq 0)$ to the state $(k+1)$, and with a service completion there is a transition from the state j to the state $(j-1)(j>0)$, the state denoting the number in the system. Before discussing queueing models in terms of birth-death equations, we take up the simplest queueing system $M/M/1$ through a simple alternative approach. This approach is based on the *rate-equality principle*, which holds for systems in steady state. The principle is stated as follows.

Rate-Equality Principle. *The rate at which a process enters a state $n(\geq 0)$ equals the rate at which the process leaves that state n. In other words, the rate of entering and the rate of leaving a particular state are the same for every state—that is,*

$$\textit{rate in} = \textit{rate out or rate up} = \textit{rate down, for every state.}$$

We have already discussed this principle under Remark (3) following Eq. (1.3.23) in Chapter 1.

3.2 The Simple $M/M/1$ Queue

In such a queueing system, the arrivals occur from an infinite source in accordance with a Poisson process with parameter λ—that is, the interarrival times are independent exponential with mean $1/\lambda$; the service times are

independently and exponentially distributed with parameter (say, μ); and there is only one server. The queue discipline is FCFS; the utilization factor is $\rho = \lambda/\mu = a, \lambda$ and μ being the arrival and service rates, respectively.

3.2.1 Steady-state solution of M/M/1

Assume that steady state exists and let

$$p_n = \lim_{t\to\infty} Pr\{N(t) = n\}, \quad n = 0, 1, 2, \ldots, \tag{3.2.1}$$

$N(t)$ being the number in the system (in the service channel and in queue, if any) at instant t; p_n is also the proportion of time the process is in state n.

We proceed to derive the equations involving p_n by using the rate-equality principle; then we proceed to solve the equations to find p_n.

Consider state n ($n \geq 0$). The system can go to the next state $(n+1)$ at rate λp_n, and it can come down from state $(n+1)$ to the original state n at rate μp_{n+1}.

For equilibrium these two rates—that is, the rate up from a particular state n to the next state $(n+1)$—and the rate down—that is, from the state $(n+1)$ to the original state n—must be equal. (In equilibrium, rate up = rate down.) This implies that

$$\begin{aligned} \lambda p_n &= \mu p_{n+1} \quad (n \geq 0) \\ \text{or} \quad p_{n+1} &= \frac{\lambda}{\mu} p_n = a p_n = a^2 p_{n-1} \\ &\cdots \quad \cdots \\ &= a^{n+1} p_0 \\ \text{or} \quad p_n &= a^n p_0, \quad n \geq 0 \end{aligned} \tag{3.2.2}$$

Using $\sum_{n=0}^{\infty} p_n = 1$, one gets, for $a < 1$,

$$p_n = (1-a)a^n, \quad n = 0, 1, 2, \ldots$$

Since $a = \rho$, we get

$$\begin{aligned} p_0 &= (1-a) = 1 - \rho \\ p_n &= (1-\rho)\rho^n, \quad n = 1, 2, \ldots \end{aligned} \tag{3.2.3}$$

The distribution is geometric and is memoryless.

3.2.1.1 Some performance measures

Let N be the number and W the waiting time in the system in steady state. We have

$$E\{N\} = \sum_{n=0}^{\infty} np_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n$$

$$= \rho(1-\rho)\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho(1-\rho)}{(1-\rho)^2} = \frac{\rho}{1-\rho} \tag{3.2.4}$$

and

$$E\{N^2\} = \sum_{n=0}^{\infty} n^2 p_n = \sum_{n=1}^{\infty} n^2(1-\rho)\rho^n$$

$$= (1-\rho)\sum\{(n^2-n)+n\}\rho^n$$

$$= (1-\rho)\frac{2\rho^2}{(1-\rho)^3} + \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{2\rho^2}{(1-\rho)^2} + \frac{\rho}{1-\rho}$$

$$= \frac{\rho+\rho^2}{(1-\rho)^2}$$

so that $\operatorname{var}\{N\} = E\{N^2\} - [E\{N\}]^2$

$$= \frac{\rho}{(1-\rho)^2}. \tag{3.2.5}$$

Using Little's formula $L = \lambda W$, we get that the expected waiting time in the system, $E\{W\}$, equals

$$E\{W\} = \frac{E\{N\}}{\lambda} = \frac{1}{\lambda}\frac{\rho}{(1-\rho)} = \frac{1}{\mu(1-\rho)} \tag{3.2.6}$$

Notes:

(1) The condition $\lambda/\mu = \rho < 1$ is necessary to get a solution. If $\lambda/\mu > 1$, then $\sum_n(\lambda/\mu)^n$ increases without limit and $p_n = 0$ for $n = 0, 1, 2, \ldots$. Thus, $\rho < 1$ is a necessary condition for existence of steady states.

(2) When $\rho \to 1$ from below, $E\{N\}$ and $\operatorname{var}\{N\}$ tend to ∞; also $p_0 \to 0$, $p_n \to 0, n \geq 1$, so that the probability that the system is empty or contains a certain fixed number is very, very small. Large variance of the number in the system implies that a randomly observed system size is likely to be very different from the expected system size.

(3) The distribution of the number N_Q in the queue can be obtained from that of N. One can thus get $E\{N_Q\}$ and $\text{var}\{N_Q\}$. We have

$$E\{N_Q\} = E\{N\} - \rho = \frac{\rho^2}{1-\rho} \quad \text{and} \quad E\{W_Q\} = \frac{E\{N_Q\}}{\lambda} = \frac{\rho}{\mu(1-\rho)}$$

3.2.2 Waiting-time distributions

We may consider two types of waiting times: (1) waiting time W_q in the queue or *queueing time* and (2) waiting time W in the system, which includes queueing time plus service time—that is, total time spent in the system by the test unit (also called *sojourn time* or *response time*). To find these distributions, queue discipline has to be taken into account. Assume that it is FCFS.

We consider at first the distribution of the waiting time or queueing time W_q of a test unit. We have $W_q = 0$ when there is no such unit in the system just before arrival of the test unit; the probability of this event is $p_0 = 1 - \rho$. Thus, $Pr\{W_q > 0\} = 1 - p_0 = \rho$—that is, the probability that an arrival has to wait in queue is ρ; this probability is large for large ρ. If the test unit finds, on arrival $n(>1)$ units in the system, then $W_q = S_n$, where

$$S_n = v_1' + v_2 + \cdots + v_n,$$

v_1' being the residual service time of the customer being served, at the epoch of his arrival and $v_2, \ldots, v_n$ being the service times of the $(n-1)$ units in the queue. The residual service time v_1' of the exponential service-time distribution with mean $1/\mu$ has an exponential distribution with the same mean $(1/\mu)$. The RVs $v_2, \ldots, v_n$ are independent exponential with mean $1/\mu$. Thus, S_n is the sum of n identically and independently distributed exponential RVs, each with mean $(1/\mu)$. The RV S_n therefore has a gamma distribution having PDF

$$\frac{\mu^n x^{n-1} e^{-\mu x}}{\Gamma(n)}, \quad x > 0. \tag{3.2.7}$$

Denote

$$w_q(x)dx = P\{x \le W_q < x + dx\}.$$

Conditioning on the number of units that the test unit finds on arrival, we get

$$\begin{aligned} w_q(x)dx &= \sum_{n=1}^{\infty} Pr\{x \le W_q < x + dx \mid \text{the test unit finds } n \text{ in the system}\} \\ &\quad \times Pr\{\text{the test unit finds } n \text{ in the system}\} \\ &= \sum_{n=1}^{\infty} \frac{\mu^n x^{n-1} e^{-\mu x}}{\Gamma(n)} dx(a_n). \end{aligned}$$

Now since $a_n = p_n$ for the Poisson arrival process, we get

$$w_q(x) = \mu e^{-\mu x}(1-\rho)\rho \sum_{n=1}^{\infty} \frac{(\mu\rho x)^{n-1}}{(n-1)!}$$

$$= \mu\rho(1-\rho)e^{-\mu(1-\rho)x}, \quad x > 0.$$

Thus, the PDF of W_q is given by

$$w_q(x) = p_0 = 1-\rho, \quad x = 0$$

$$= \mu\rho(1-\rho)e^{-\mu(1-\rho)x}, \quad x > 0, \tag{3.2.8}$$

which is a modified exponential distribution.

The LST of W_q can be obtained from (3.2.8).

We can use the same argument to obtain directly the LST of waiting-time distribution. Let $w_q^*(s)$ be the LST of the waiting time W_q in the queue and $w_q^*(s \mid n)$ be the conditional LST of the waiting time in the queue given that the test unit finds n in the system on his arrival. We have

$$w_q^*(s) = \sum_{n=0}^{\infty} a_n w_q^*(s \mid n),$$

where a_n is the probability that the arrival finds n in the system. For Poisson arrivals, however, $a_n = p_n = (1-\rho)\rho^n$. If he finds $n = 0$ customers, he does not wait, if he finds $n(\geq 1)$, he waits until the services of n are completed. The service time of n being equal to the sum of n IID exponential variables with mean $1/\mu$, we get

$$w_q^*(s \mid n) = \left(\frac{\mu}{s+\mu}\right)^n, \quad n \geq 1.$$

Thus,

$$w_q^*(s) = (1-\rho) + (1-\rho)\sum_{n=1}^{\infty} \rho^n \left(\frac{\mu}{s+\mu}\right)^n \tag{3.2.9}$$

$$= (1-\rho) + (1-\rho)\frac{\rho\mu/(s+\mu)}{1-\frac{\rho\mu}{s+\mu}}$$

$$= (1-\rho) + \frac{\mu\rho(1-\rho)}{s+\mu(1-\rho)}. \tag{3.2.9a}$$

The LST of the waiting time in the system W (or sojourn time) is given by

$$w^*(s) = w_q^*(s)\frac{\mu}{s+\mu}$$

$$= \frac{\mu(1-\rho)}{s+\mu}\frac{s+\mu}{s+\mu(1-\rho)}$$

$$= \frac{\mu(1-\rho)}{s+\mu(1-\rho)}. \tag{3.2.10}$$

so that the PDF of the waiting time (in the system) is obtained as

$$w(x) = \mu(1-\rho)e^{-\mu(1-\rho)x}, \quad x \geq 0; \tag{3.2.11}$$

the distribution is exponential with parameter $\mu(1-\rho)$.

Notes:

(1) From (3.2.9), we get

$$w_q(x) = Pr\{W_q \leq x\} = (1-\rho)\delta(x) + (1-\rho)\sum_{n=1}^{\infty} \rho^n \{B(x)\}^{n*} \tag{3.2.12}$$

where $\{B(x)\}^{n*}$ is the n-fold convolution of the exponential service time DF $B(x)$.

The form (3.2.9) implies that $w_q^*(s)$ is the geometric compounding of the exponential service time; so also (3.2.12) does imply the same about $W_q(x)$.

(2) Moments of W_q can be easily obtained as follows. We have

$$\begin{aligned} E[W_q] &= -\frac{d}{ds} w_q^*(s)|_{s=0} \\ &= \frac{(1-\rho)\lambda}{(s-\lambda+\mu)^2}\bigg|_{s=0} = \frac{\lambda}{\mu(\mu-\lambda)} \\ &= \frac{\rho}{\mu(1-\rho)} \quad \text{and} \end{aligned} \tag{3.2.13}$$

$$\begin{aligned} E[W_q^2] &= -\frac{d^2}{ds^2} w_q^*(s)|_{s=0} \\ &= \frac{2(1-\rho)\lambda}{(s-\lambda+\mu)^3}\bigg|_{s=0} \\ &= \frac{2(1-\rho)\lambda}{(\mu-\lambda)^3} = \frac{2\lambda}{\mu(\mu-\lambda)^2} \end{aligned} \tag{3.2.14}$$

Hence,

$$\begin{aligned} \text{var}\,[W_q] &= \frac{2\lambda}{\mu(\mu-\lambda)^2} - \left[\frac{\lambda}{\mu(\mu-\lambda)}\right]^2 \\ &= \frac{\lambda(2\mu-\lambda)}{\mu^2(\mu-\lambda)^2} = \frac{\rho(2-\rho)}{(\mu-\lambda)^2} \end{aligned} \tag{3.2.15}$$

(3) Values of $E(N)$, var(N), $E(W)$, and var(W) are given below for certain values of $\mu = 1$ and $\lambda = \rho$.

$\rho(=\lambda)$	0.4	0.6	0.8	0.9	0.98	0.99	0.995
$E(N)$	0.67	1.5	4.0	9.0	49.0	99.0	199
var(N)	1.11	3.75	20.0	90.0	2450.0	$10^2 \times 99.0$	$10^3 \times 199.0$
$E(W)$	1.67	2.50	5.0	10.0	50.0	100.0	200.0
var(W)	2.78	6.25	25.0	100.0	2500.0	$10^4 \times 1.0$	$10^4 \times 4.0$

(4) The preceding gives the unconditional distribution of the queueing time of a test unit. From (3.2.9a) the LST of the conditional distribution of waiting time given that there is a positive wait (that is, a test unit has to wait) is given of

$$cw_q^*(s) = \frac{\mu\rho(1-\rho)}{\rho[s+\mu(1-\rho)]} = \frac{\mu(1-\rho)}{s+\mu(1-\rho)}.$$

The conditional distribution of the queueing time given that the test unit has to wait, has the PDF $f_c(x)$ given by

$$\begin{aligned} f_c(x)dx &= P\{x \le W_q < x + dx \mid \text{test unit has to wait}\} \\ &= \frac{\mu\rho(1-\rho)e^{-\mu(1-\rho)x}}{\rho}\,dx \\ &= \mu(1-\rho)e^{-\mu(1-p)x}\,dx, \quad x > 0. \end{aligned} \tag{3.2.16}$$

The conditional distribution is exponential with mean $1/\mu(1-\rho)$.

(5) Service in Random Order (SIRO).

It is found that for many switching systems, SIRO (Service in Random Order) discipline gives a more realistic approximation.

It can be seen that queue-length distributions under both the FCFS and SIRO disciplines will be the same, and so also will be the expected queue length and the expected queueing time. However, the distributions of queueing time for different disciplines will be different. Starting from Vaulot (1946), Riordon (1953), Kingman (1962), and Cohen (1982) also investigated $M/M/1$ queue with SIRO discipline. Recently, Flatto (1997) finds the queueing time distribution with SIRO discipline in an explicit form. He shows that, as $t \to \infty$

$$P\{W_Q > t\} \sim \alpha t^{-5/6} \exp\{-\beta t - \gamma t^{1/3}\}$$

where α, β, γ are expressed in terms of the traffic intensity ρ.

He also finds the LST of W_Q.

3.2.3 The output process

The problem of output (efflux, departures) from a queueing system was first considered by Morse (1955). He observed that the output of a Poisson input queue with a single channel having exponential service time and in steady state must be Poisson with the same rate as the input. A formal proof was first given by Burke (1956). The interesting result is considered next.

Theorem 3.1. *In an $M/M/1$ queueing system in steady state, the interdeparture times are independently and identically distributed exponential random variables with mean $1/\lambda$, where λ is the parameter of the input (Poisson) process. In other words, the output process is Poisson with the same parameter as the input process.*

Proof: Let $N(t)$ be the number in the system at time t and $t_1', t_2', \ldots, t_n', t_{n+1}' \ldots$ denote the successive departure instants so that $L = t_{n+1}' - t_n'$ is the nth interdeparture interval or period. Let

$$F_k(t) = P\{N(t_n' + t) = k, \quad t_{n+1}' - t_n' > t\}, \quad t > 0, \ k = 0, 1, \ldots, \tag{3.2.17}$$

be the joint probability distribution of $N(t)$ and $t_{n+1}' - t_n'$. Since the probability that a departing customer leaves k in the system is equal to the probability that the number in the system is k for Poisson input (that is, $d_k = p_k$), we have

$$F_k(0) = p_k = (1 - \rho)\rho^k, \quad k = 0, 1, 2, \ldots \tag{3.2.18}$$

For an infinitesimal interval of length dt

$$F_0(t + dt) = F_0(t)(1 - \lambda\, dt) + o(dt) \tag{3.2.19}$$

and

$$F_k(t + dt) = F_k(t)[1 - \lambda\, dt - \mu\, dt] + F_{k-1}(t)\lambda\, dt + o(dt).$$

These equations reduce to

$$\begin{aligned} F_0'(t) &= -\lambda F_o(t) \quad \text{and} \\ F_k'(t) &= \lambda F_{k-1}(t) - (\lambda + \mu)F_k(t), \quad k = 1, 2, 3, \ldots \end{aligned} \tag{3.2.20}$$

Subject to the initial condition $F_k(0) = p_k$, the unique solution of (3.2.20) is given by

$$\begin{aligned} F_k(t) &= p_k e^{-\lambda t} \\ &= (1 - \rho)\rho^k e^{-\lambda t}. \end{aligned} \tag{3.2.21}$$

Again,

$$\begin{aligned} Pr\{N(t'_{n+1}+0) = k, \quad t \le t'_{n+1} - t'_n < t + dt\} \\ &= F_{k+1}(t) \cdot Pr\{\text{one service completion in } (t, t+dt)\} \\ &= F_{k+1}(t)[\mu\, dt + o(dt)] \\ &= (1-\rho)\rho^{k+1} e^{-\lambda t} \mu\, dt + o(dt) \\ &= (1-\rho)\rho^k \lambda e^{-\lambda t}\, dt + o(dt), \end{aligned} \tag{3.2.22}$$

which proves the independence of $N(t'_{n+1}+0)$ and $(t'_{n+1} - t'_n)$. It follows that $L = t'_{n+1} - t'_n$ has density $\lambda e^{-\lambda t}$.

Now we look at the independence of the interdeparture intervals. Let Λ represent the set of lengths of an arbitrary number of interdeparture intervals subsequent to the interval of length L. Let $P(.)$ represent the probability function of the variable within ( ). The Markov property implies that

$$P[\Lambda \mid N(L)] = P[\Lambda \mid N(L), L] \tag{3.2.23}$$

Since $N(L)$ and L are independent,

$$P[N(L), L] = P[N(L)]P(L). \tag{3.2.24}$$

The joint probability function of the initial interval length, the state at the end of the interval, and the set of subsequent interval lengths may be expressed (using (3.2.23) and (3.2.24)) as

$$\begin{aligned} P[L, N(L), \Lambda] &= P[\Lambda \mid N(L), L]P[N(L), L] \\ &= P[\Lambda \mid N(L)]P[N(L)]P(L) \\ &= P[\Lambda, N(L)]P(L). \end{aligned}$$

Thus,

$$P(L, \Lambda) = \sum_{N(L)=0}^{\infty} P[\Lambda, N(L)]P(L) = P(L)P(\Lambda)$$

which implies the mutual independence of all intervals. This completes the proof. ■

Notes:

(1) For the existence of steady state, the condition $\lambda < \mu$ is essential. The output process for large t is Poisson, even for $\lambda \ge \mu$, when steady state does not hold. For $M/M/1$ queue, the output process for large t is Poisson with rate $\min(\lambda, \mu)$, (Goodman and Massey, 1984).

(2) That the expected inter-output (interdeparture) rate for a single-server Poisson input queue (with arrival rate λ) is λ and can also be shown as follows.

Conditioning on the state in which a departure leaves the system (the state $n > 0$ or $n = 0$), we have

$$\begin{aligned} E(L) &= E\{L \mid N = n(>0)\}[Pr\{N = n(>0)\}] + E(L \mid N = 0)\, Pr\{N = 0\} \\ &= (1/\mu)(1 - p_0) + (1/\lambda + 1/\mu)\, p_0 \\ &= (1/\mu) + (1/\lambda)(1 - \rho) \\ &= 1/\lambda. \end{aligned}$$

The result holds for a $GI/G/1$ queue (see Section 7.4).

(3) Robustness of the $M/M/1$ Queue

Albin (1984) examines the robustness or insensitivity of the $M/M/1$ queueing model to specific perturbations of the input process. Perturbations considered are deviations from the exponential distribution of the interarrival times and from the assumption of independence between successive interarrival times. Let θ be a parameter of the system, let $L(\theta)$ be the expected number in the system in steady state in the standard $M/M/1$ queue, and let $L(\theta + \varepsilon)$ be the expected number in the perturbed system. Then expanding by Taylor's series, we get

$$L(\theta + \varepsilon) \simeq L(\theta) + \varepsilon L'(\theta).$$

$L'(\theta)$ is called the perturbation rate, and the system is called sensitive when $|L'(\theta)|$ is large compared with $L(\theta)$. For example, for an $M/M/1$ system with ρ as parameter,

$$\begin{aligned} L(\rho) &= \frac{\rho}{(1 - \rho)} \\ L'(\rho) &= \frac{1}{(1 - \rho)^2} = \left[\frac{1}{\rho^2}\right][L(\rho)]^2, \end{aligned}$$

which shows that the expected number in the $M/M/1$ system is sensitive to small changes in ρ—in particular, for ρ near to 1. Her results in the specific cases examined indicate that the expected number in the system is robust or insensitive to specific perturbations in the arrival process, such as perturbations by insertion of a few short interarrival times, of an occasional batch arrival, or of small dependencies between successive interarrival times. See Zolotarev (1977), Rachev (1989), and Kalashnikov and Rachev (1990) for books on stability of queues.

3.2.4 Semi-Markov process analysis

We shall now consider the approach through the semi-Markov process (discussed in Section 1.9) for this system. Here we consider that a transition occurs with the arrival or departure of a unit—that is, t_n $(n = 0, 1, 2, \ldots)$ is the epoch at which the nth transition (through an arrival or service completion) occurs. With the notation of Section 1.9, $\{Y(t), t \geq 0\}$ is a semi-Markov process having $\{X_n, n \geq 0\}$ [where $X_n = N(t_n + 0), Y(t) = X_n, t_n \leq t < t_{n+1}$] as its embedded Markov chain. If u and v denote the interarrival and service times, respectively, then $Pr\{u < v\} = \lambda/(\lambda + \mu), Pr\{u > v\} = \mu/(\lambda + \mu)$, and $\{\min(u, v)\}$ is exponential with parameter $(\lambda + \mu)$. (See Problem 1.17.)

We have

$$\begin{aligned}
Q_{0,1}(t) &= Pr\{\text{transition occurs from state 0 to state 1 by time } t\} \\
&= Pr\{\text{an arrival occurs by time } t\} \\
&= 1 - e^{-\lambda t} \\
Q_{j,j+1}(t) &= Pr\{\text{transition occurs from state } j \text{ to state } j+1 \text{ by time } t\} \\
&= Pr\{\text{one transition (arrival or service completion) occurs by time } t\} \\
&\quad \times Pr\{\text{transition is through an arrival}\} \\
&= Pr\{\min(u, v) \leq t\} Pr\{u < v\} \\
&= \left[1 - e^{-(\lambda+\mu)t}\right] \frac{\lambda}{\lambda + \mu}, \quad j > 0.
\end{aligned}$$

Similarly,

$$Q_{j,j-1}(t) = \left[1 - e^{-(\lambda+\mu)t}\right]\left[\frac{\mu}{\lambda + \mu}\right]; \quad j > 0.$$

Thus,

$$p_{i,j} = \lim_{t\to\infty} Q_{i,j}(t) \quad \text{gives}$$

$$p_{0,1} = 1, \quad p_{j,j+1} = \frac{\lambda}{\lambda + \mu}, \quad p_{j,j-1} = \frac{\mu}{\lambda + \mu}, \quad j \geq 1$$

and

$$p_{i,j} = 0 \text{ in all other cases.}$$

Thus, $v_k = \lim_{n\to\infty} p_{i,j}^{(n)}$ are given as the unique solution of $\boldsymbol{V} = \boldsymbol{V}P$—that is,

$$v_j = \sum_k v_k p_{kj}.$$

We get

$$v_0 = v_1 p_{1,0} = \left(\frac{\mu}{\lambda+\mu}\right) v_1$$
$$v_1 = v_0 p_{0,1} + v_2 p_{2,1}$$
$$= v_0 + v_2 \left(\frac{\mu}{\lambda+\mu}\right).$$

and for $j > 1$

$$v_j = v_{j-1} p_{j-1,j} + v_{j+1} p_{j+1,j}$$
$$= v_{j-1}\left(\frac{\lambda}{\lambda+\mu}\right) + v_{j+1}\left(\frac{\mu}{\lambda+\mu}\right).$$

Solving the above difference equation, we get

$$v_j = A + B(\lambda/\mu)^j, \quad j = 1, 2, \ldots,$$

where A, B are constants. Evaluating the constants with the help of expressions for v_0, v_1, we get

$$v_j = \left(\frac{\lambda+\mu}{\lambda}\right)\left(\frac{\lambda}{\mu}\right)^j v_0, \quad j = 1, 2, \ldots,$$

From Equation (1.9.4) of Chapter 1, we get

$$p_k = \lim_{t\to\infty} Pr\{Y(t) = k\}$$
$$= \frac{v_k \mu_k}{\sum v_j \mu_j}, \quad k = 0, 1, \ldots,$$

where μ_k= expected sojourn time in state k. We get

$$\mu_0 = E\{u\} = 1/\lambda$$
$$\mu_k = \text{expected time for a transition}$$
$$= E\{\min(u, v)\} = \frac{1}{\lambda+\mu}, \quad j \geq 1$$

and

$$v_j \mu_j = \frac{1}{\lambda}\left(\frac{\lambda}{\mu}\right)^j v_0$$

so that

$$\sum_{j=0}^{\infty} v_j \mu_j = (v_0/\lambda)\frac{1}{1-\lambda/\mu}$$

$$= \frac{\mu}{\lambda}\frac{v_0}{(\mu-\lambda)}.$$

Thus,

$$p_k = \frac{(1/\lambda)(\lambda/\mu)^k v_0}{(\mu/\lambda)\{1/(\mu-\lambda)\}v_0}$$

$$= (1-\lambda/\mu)(\lambda/\mu)^k$$

$$= (1-\rho)\rho^k, \quad k = 0, 1, 2, \ldots.$$

Notes:

(1) Here the system size $N(t)$ is semi-Markovian (as well as Markovian).

(2) $v_0 < p_0$ and $v_k > p_k, k = 1, 2, \ldots$

3.3 System with Limited Waiting Space: The $M/M/1/K$ Model

For the simple queue $M/M/1$, the assumption is that the system can accommodate any number of units. In this model, we assume that the system can accommodate a finite number of units—say, K—including the one being served, if any. Customers arrive in accordance with a Poisson process with rate—say, λ; a customer will join the system whenever he finds less than K in the system and a customer who arrives when there are K in the system leaves the system and is lost to the system. Service time is exponential with rate μ. A queue with limited waiting space is known as a queue with *finite buffer.*

3.3.1 Steady-state solution

Here the system will behave as an ordinary $M/M/1$ queue so long as the number in the system is less than K. Then when it reaches the state K or the system contains K customers, no more arrival is allowed to the system, and the number in the system cannot exceed K. From state K, only departure is possible. Using the rate-equality (rate up = rate down) principle, the balance

equations can be written as follows:

$$\lambda p_n = \mu p_{n+1}, \quad n = 0, 1, 2, \ldots, K-1 \tag{3.3.1}$$

Solving the equations recursively, we get

$$p_n = p_0 a^n, \quad a = \lambda/\mu, \quad n = 0, 1, 2, \ldots, K. \tag{3.3.2}$$

Using the normalizing condition

$$\sum_{n=0}^{K} p_n = 1, \quad \text{we get } p_0 \sum_{n=0}^{K} a^n = 1$$

so that

$$p_0 = \left[\sum_{n=0}^{K} a^n\right]^{-1} = \frac{1-a}{1-a^{K+1}}, \quad \lambda \neq \mu$$

$$= \frac{1}{K+1}, \quad \lambda = \mu.$$

Thus, we have, for $n = 0, 1, \ldots, K$,

$$p_n = p_0 a^n = \frac{(1-a)a^n}{1-a^{K+1}}, \quad \lambda \neq \mu$$

$$= \frac{1}{K+1}, \quad \lambda = \mu. \tag{3.3.3}$$

Its PGF is given by

$$G(s) = \sum_{n=0}^{K} p_n s^n = \frac{1-a}{1-a^{K+1}} \left[\frac{1-(as)^{K+1}}{1-as}\right] \quad \text{for } a \neq 1. \tag{3.3.4}$$

The distribution of the number in the system is uniform for $a = 1$ and truncated geometric for $a \neq 1$.

3.3.2 Expected number in the system L_K

We have for $\lambda = \mu$

$$L_K = \sum_{n=0}^{K} n p_n$$

$$= \sum_{n=0}^{K} \frac{n}{K+1} = \frac{K}{2};$$

for $\lambda \neq \mu$,

$$\begin{aligned} L_K &= \frac{(1-a)a}{1-a^{K+1}} \sum_{n=0}^{K} n a^{n-1} \\ &= \frac{(1-a)a}{1-a^{K+1}} \sum_{n=0}^{K} \frac{d}{da}(a^n) \\ &= \frac{(1-a)a}{1-a^{K+1}} \frac{1-(K+1)a^K + Ka^{K+1}}{(1-a)^2} \\ &= \frac{a}{1-a} - \frac{(K+1)a^{K+1}}{1-a^{K+1}}. \end{aligned} \tag{3.3.5}$$

Remarks:

(1) It may be noted that the finite series $\sum_{n=0}^{K} a^n$ has a sum for all values of a; thus, the steady-state solution exists for all values of a.

(2) Assuming $a < 1$ and taking limit as $K \to \infty$, we get the corresponding results for an $M/M/1$ model.

(3) The system with $K = 1$ (with no waiting room at all, as it may happen in a telephone kiosk) is known as a system with *blocked calls cleared* with a single server.

(4) The effective input rate λ' to the system $M/M/1/K$ is the expected number of customers joining the system in unit time and equals

$$\begin{aligned} \lambda' = \lambda(1 - p_K) &= \frac{\lambda(1-a^K)}{1-a^{K+1}}, \quad \lambda \neq \mu, \\ &= \frac{\lambda K}{K+1}. \quad \lambda = \mu. \end{aligned}$$

The average interarrival time (to the system) is $1/\lambda'$ (which is $> 1/\lambda$).

(5) The average number of customers diverted from the system or lost to the system in unit time equals

$$\begin{aligned} \lambda - \lambda(1 - a_K) = \lambda p_K &= \frac{\lambda(1-a)a^K}{1-a^{K+1}}, \quad a \neq 1, \\ &= \frac{\lambda}{K+1}, \quad a = 1. \end{aligned}$$

(6) The utilization factor of the service station is not equal to a as in usual single-server models, but equals

$$\rho' = \frac{\lambda'}{\mu} = a(1 - p_K) = \frac{a(1 - a^K)}{1 - a^{K+1}}, \quad \lambda \neq \mu,$$
$$= aK/(K + 1), \quad \lambda = \mu.$$

(7) The expected number of customers leaving the service station (after being served) in unit time equals

$$\mu\rho' = \frac{\lambda(1 - a^K)}{1 - a^{K+1}}, \quad \lambda \neq \mu,$$
$$= \frac{\lambda K}{K + 1}, \quad \lambda = \mu.$$

(8) Naor (1969) has used the model $M/M/1/K$ to study the regulation of queue size by levying tolls. (See Problems and Complements 3.4.) Rue and Rosenshine (1981) have extended Naor's arguments to obtain a policy for "individual optimum" in case of M classes of customers. Bounds on this policy and a numerical example with three classes of customers are also considered.

(9) The mean waiting time in queue equals L_K/λ'.

(10) Though the *arrival* process is Poisson, the *input* for this truncated system is not truly Poisson. PASTA does not hold in this case. The probability a_n that an effective arrival (an entry to the system) finds n in the system [in an infinitesimal interval $(t, t + h)$] can be obtained by applying Baye's theorem. We have

$$a_n = Pr\{n \text{ in system} \mid \text{an arrival is about to occur}\}$$
$$= \frac{Pr\{\text{an arrival is about to occur} \mid n \text{ in system}\} \times p_n}{\sum_{k=0}^{K-1} Pr\{\text{an arrival is about to occur} \mid k \text{ in system}\}}$$
$$= \lim_{h \to 0} \frac{\{\lambda h + o(h)\} p_n}{\sum_{k=0}^{K-1} \{\lambda h + o(h)\} p_k}$$
$$= \frac{\lambda p_n}{\sum_{k=0}^{K-1} \lambda p_k} \tag{3.3.6}$$
$$= \frac{p_n}{1 - p_K}, \quad n = 0, 1, \ldots, K - 1. \tag{3.3.7}$$

3.3.3 Equivalence of an $M/M/1/K$ model with a two-stage cyclic model

Consider a cyclic model with a fixed number K of jobs circulating endlessly between two servers I and II. The two servers have independent exponential service-time distributions with rates μ and λ, respectively, the order of service at

each of the service counters being FCFS. As long as the number of customers at server I is less than K, customers will arrive there from server II, the interarrival times being distributed as the service time of server II—that is, exponential with rate λ. When all the customers are at server I, no further arrival can take place. Thus, the model is the same as the $M/M/1/K$ model; the event that there are n customers either in queue or in service with server I and $K-n$ customers with server II in the cyclic model is the same as that of n customers in an $M/M/1/K$ model, and so, for $n = 0, 1, 2, \ldots, K$,

$$\begin{aligned} p(n, K-n) &= Pr(n \text{ customers with server I and } K-n \text{ with server II}) \\ &= p_n, \quad \text{as given by (3.3.3).} \end{aligned}$$

The utilization factors of servers I and II are given by

$$\rho_1 = 1 - p_0 \quad \text{and} \quad \rho_2 = 1 - p_K, \quad \frac{\rho_1}{\rho_2} = \rho.$$

The expected numbers N_1 and N_2 with the subsystems (servers I and II) are given by

$$E(N_1) = \sum_{n=0}^{K} np_n$$

and

$$E(N_2) = K - E(N_1).$$

The average time taken for a customer to go through *both* the servers to make a complete cycle is

$$\frac{K}{(\rho_1 \mu)} = \frac{K}{(\rho_2 \lambda)}.$$

The preceding model is useful for computer systems where the CPU can be taken as server I and I/O unit as server II. (See Problems and Complements 5.6 for the model with general service time.)

3.4 Birth-and-Death Processes: Exponential Models

Consider a queueing system where the arrivals occur from an infinite source in accordance with a Poisson process and the service times are independently and exponentially distributed. The rates of arrival and service are state-dependent: when there are n in the system, the rate of arrival is λ_n and the rate of service is μ_n. The queue discipline is FCFS, and there is no restriction as to the number

of servers.

$$p_n = \lim_{t\to\infty} P\{N(t) = n\}, \quad n = 0, 1, 2, \ldots$$

We shall obtain the balance equations by using the rate-equality principle.

Consider state 0. The process can leave or go out from state 0 only when there is an arrival, and then the change occurs from state 0 to state 1. It can enter the state 0 from state 1 only with a departure. Using the rate-equality principle, we get

$$\lambda_0 p_0 = \mu_1 p_1. \tag{3.4.1}$$

Consider an arbitrary state n ($n = 1, 2, \ldots$). The process can leave state n in two ways, either through an arrival (when the state changes from n to $n+1$) or through a departure (when the state changes from n to $(n-1)$). The rate at which the process leaves state n is thus $(\lambda_n + \mu_n)$; the proportion of time it is in state n is p_n, and the total rate at which the process leaves state n is $(\lambda_n + \mu_n) p_n$. On the other hand, the process can enter the state either from state $(n-1)$ through an arrival or from state $(n+1)$ through a departure. The proportion of time the process is in state $(n-1)$ is p_{n-1}, and the rate at which it can enter state n through an arrival from state $(n-1)$ is λ_{n-1}. Thus, the rate of entering state n from state $(n-1)$ through an arrival is $\lambda_{n-1} p_{n-1}$; similarly the rate of entering state n from state $(n+1)$ through a departure is $\mu_{n+1} p_{n+1}$. Thus, we have (by rate up = rate down principle)

$$\lambda_n p_n = \mu_{n+1} p_{n+1}, \quad n = 0, 1, 2, \ldots \tag{3.4.2}$$

Thus,

$$\begin{aligned} p_{n+1} &= \frac{\lambda_n}{\mu_{n+1}} p_n \\ &= \frac{\lambda_n}{\mu_{n+1}} \frac{\lambda_{n-1}}{\mu_n} p_{n-1} \\ &\cdots \\ &\cdots \\ &= \prod_{k=0}^{n} \frac{\lambda_k}{\mu_{k+1}} p_0, \quad n = 0, 1, 2, \ldots, \quad \text{or} \end{aligned}$$

$$p_n = \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}} p_0, \quad n = 1, 2, \ldots \tag{3.4.3}$$

Using $\sum_{n=0}^{\infty} p_n = 1$, we get

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}}. \tag{3.4.4}$$

The necessary and sufficient condition for the existence of a steady state is the convergence of the infinite series $\sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \lambda_k/\mu_{k+1}$, which occurs in the denominator of (3.4.4). When the series converges, p_0 can be obtained by using (3.4.4).

The preceding birth-death process model can be used to study various queueing systems. For example, the $M/M/1$ model considered in Section 3.2 is the model with constant arrival and service rates

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \ldots, \quad \text{and}$$
$$\mu_n = \mu, \quad n = 0, 1, 2, \ldots$$

Using (3.4.4) and (3.4.3) with these values of λ_n, μ_n, we at once get the steady-state results. The $M/M/1/K$ queueing model can be obtained by putting

$$\lambda_n = \lambda, \quad n = 0, 1, \ldots, K-1$$
$$= 0, \quad n = K,$$
$$\mu_n = \mu, \quad n = 1, 2, \ldots, K.$$

3.5 The $M/M/\infty$ Model: Exponential Model with an Infinite Number of Servers

We consider here the exponential model with an infinite number of servers. An example of where such a model can be appropriate is a service facility with provision of self-service. It is a birth-death model with

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \ldots, \quad \text{and}$$
$$\mu_n = n\mu, \quad n = 1, 2, \ldots$$

The solution is given by

$$p_n = \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}} = p_0 \prod_{k=0}^{n-1} \frac{\lambda}{(k+1)\mu}$$
$$= p_0 \frac{\lambda^n}{\mu(2\mu)\ldots(n\mu)} = p_0 \frac{(\lambda/\mu)^n}{n!}, \quad n = 0, 1, 2, \ldots$$

To find p_0, we use

$$1 = \sum_{n=0}^{\infty} p_n = \left[\sum_{n=0}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \right] p_0$$
$$= e^{\lambda/\mu} p_0$$

so that

$$p_0 = e^{-\lambda/\mu}$$

and, thus,

$$p_n = \frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}, \quad n = 0, 1, 2, \ldots \tag{3.5.1}$$

The distribution is Poisson with mean λ/μ. The expected number of customers in the system is λ/μ, and the expected response time is $1/\mu = [(\lambda/\mu)/\lambda]$, the average time required for service (as is obvious). The result holds irrespective of the magnitude of λ/μ.

Notes:

(1) The result holds irrespective of the form of the service time distribution as well (as will be shown in Section 6.10.1).

(2) Interpreting W as the mean time of service of a unit, the mean number of units in the system as L, it is seen that $L = \lambda W$ is satisfied for this model.

(3) Newell (1984) discusses in his book several interesting aspects of this model.

3.6 The Model *M/M/c*

3.6.1 Steady-state distribution

We consider a queue with Poisson input (having rate, say, λ) and with c ($1 \leq c \leq \infty$) parallel service channels having IID exponential service time distribution, each with rate—say, μ. If there are n units in the system, and n is less than c, then, in all, n channels are busy and the interval between two consecutive service completions (being the minimum of n IID exponential RVs each with parameter μ) is again exponential with rate $n\mu$ (see Problem 1.17(b)). If there are $n(\geq c)$ in the system, then all the c channels are busy and the interval between two consecutive service completions is exponential with rate $c\mu$. Thus, we have a birth-death model having constant arrival (birth) rate λ and state-dependent service (death) rate

$$\begin{aligned} \mu_n &= n\mu, \quad n = 0, 1, 2, \ldots, c \\ &= c\mu, \quad n = c+1, c+2, \ldots \end{aligned}$$

Denote $\rho = \lambda/c\mu$. Assume that steady state exists and that the system is in steady state. Putting the values of λ_n and μ_n in (3.4.3) and (3.4.4), we get, for

$n = 1, 2, \ldots, c$,

$$p_n = \frac{\lambda\lambda \ldots \lambda}{(\mu)(2\mu)\ldots(n\mu)} p_0 = \frac{(\lambda/\mu)^n}{n!} p_0,$$
$$= \frac{\lambda}{n\mu} p_{n-1} \tag{3.6.1}$$

and for $n = c, c+1, c+2, \ldots$,

$$p_n = \frac{(\lambda)(\lambda)\ldots(\text{to } n \text{ factors})}{[(\mu)(2\mu)\ldots(c\mu)][(c\mu)(c\mu)\ldots(\text{to } (n-c) \text{ factors})]} p_0$$
$$= \frac{\lambda^n}{c!\mu^c c^{n-c}\mu^{n-c}} p_0 = \frac{(\lambda/\mu)^n}{c!c^{n-c}} p_0$$
$$= \frac{\lambda}{c\mu} p_{n-1} = \rho^{n-c} p_c. \tag{3.6.2}$$

Equations (3.6.1) and (3.6.2) can be written also as

$$[\min(n, c)]\mu p_n = \lambda p_{n-1}, \quad n = 1, 2, \ldots$$

The normalizing condition $\sum_{n=0}^{\infty} p_n = 1$ yields

$$p_0^{-1} = 1 + \sum_{n=1}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \sum_{n=c}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}} = \sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{1}{c!c^{-c}} \sum_{n=c}^{\infty} \left(\frac{\lambda}{c\mu}\right)^n. \tag{3.6.3}$$

In order that steady-state solutions exist, the series $\sum_{n=c}^{\infty} (\lambda/c\mu)^n$ must be convergent; this, in turn, implies that ρ is less than 1. Thus, when $\rho < 1$

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1 - \lambda/c\mu)}\right]^{-1} \tag{3.6.3a}$$

and the steady-state distribution is given by (3.6.1) and (3.6.2) with p_0 given by (3.6.3a).

Notes:

(1) The steady-state probabilities p_n satisfy the recurrence relations:

$$p_n = \frac{\lambda}{n\mu} p_{n-1} = \frac{c}{n}\rho p_{n-1}, \quad n = 1, 2, \ldots, c-1$$
$$= \frac{\lambda}{c\mu} p_{n-1} = \rho p_{n-1}, \quad n = c, c+1 \ldots$$

Thus, for $n < c$, $p_n/p_{n-1} > 1$ if $(n/c) < \rho < 1$; in this case p_n is monotone increasing in n until n exceeds $c\rho$, then is monotone decreasing until $n = c$.

For $n > c$, p_n is monotone decreasing in n. The mode of the distribution is n for which $n < c\rho \leq n + 1$, and also $n + 1$, when $c\rho$ is an integer.

(2) For finite n, $\{p_n\}$ behaves like a Poisson distribution for $n \leq c$ and like a geometric distribution for $n > c$.

(3) The probability that an arriving unit has to wait on arrival is given by

$$C \equiv C\left(c, \frac{\lambda}{\mu}\right) = Pr(N \geq c) = \sum_{n=c}^{\infty} p_n$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(1-\rho)} p_0 = \frac{p_c}{1-\rho}. \tag{3.6.4}$$

This is known as Erlang's C formula (second formula) (or Erlang delay probability). Tables for different values of c and (λ/μ) are available (Descloux, 1962).

That is, $P\{W_\alpha > 0\} = C(c, a), a = \lambda/\mu$

$C(c, a)$ can also be interpreted as the fraction of time all the c servers are busy (by virtue of PASTA).

When $c = 1, C(c, a) = a = \rho$.

(4) Let Q be the number in queue. Then for $j = 0, 1, 2, \ldots$

$$P\{Q = j, \ W_Q > 0\} = P\{N = c + j\}$$
$$= \rho^j p_c \quad \text{(from (3.6.2))}.$$
$$= \rho^j(1-\rho)C(c, a) \quad \text{(from (3.6.4))}$$

Thus

$$P\{Q = j \mid W_Q > 0\}$$
$$= \frac{P\{Q = j, \ W_Q > 0\}}{P\{W_Q > 0\}}$$
$$= (1-\rho)\rho^j, \quad j = 0, 1, 2, \ldots$$

The conditional distribution is geometric with mean $\rho/(1-\rho)$. It follows that

$$E\{Q\} = \frac{\rho}{1-\rho} C(c, a).$$

This was obtained directly in Section 3.6.2.1.

(5) The state probabilities of an $M/M/c$ system in steady state can be expressed also as

$$\begin{aligned} P\{N=n\} &= p_n \\ &= c!a^{-c}(1-\rho)C(c,a), \quad n=0 \\ &= \frac{a^n}{n!}p_0, \quad n=1,2,\ldots,c-1 \\ &= \frac{a^n}{c!c^{n-c}}p_0 = \rho^{n-c}(1-\rho)C(c,a), \quad n=c,c+1,\ldots \end{aligned}$$

(6) The $M/M/\infty$ results are useful as an approximation for an $M/M/c$ queue when $\rho < 1$.

3.6.2 Expected number of busy and idle servers

The expected number of busy servers $E(B)$ is given by

$$\begin{aligned} E(B) &= \sum_{n=0}^{c-1} np_n + \sum_{n=c}^{\infty} cp_n \\ &= \frac{\lambda}{\mu}\left[\sum_{n=1}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^{n-1}}{(n-1)!} + \frac{\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!(1-\rho)}\right]p_0 \\ &= \frac{\lambda}{\mu}\left[\sum_{m=0}^{c-2}\frac{\left(\frac{\lambda}{\mu}\right)^{m}}{m!} + \frac{\{(1-\rho)+\rho\}\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!(1-\rho)}\right]p_0 \\ &= \frac{\lambda}{\mu}\left[\sum_{m=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^{m}}{m!} + \frac{\left(\frac{\lambda}{\mu}\right)^{c}}{c!(1-\rho)}\right]p_0 \\ &= \frac{\lambda}{\mu}p_0^{-1}p_0 = \frac{\lambda}{\mu} = c\rho. \end{aligned} \tag{3.6.5}$$

Hence, the expected number of idle servers $E(I)$ is given by

$$\begin{aligned} E(I) &= E(c-B) = E(c) - E(B) \\ &= c - c\rho = c(1-\rho). \end{aligned} \tag{3.6.6}$$

Note: The results also follow from Little's Law as applied to servers.

3.6.2.1 Expected number in the system E(N)

$$E(N) = E(B) + E(Q)$$

where $E(Q)$ is the expected number in queue. We have

$$
\begin{aligned}
E(Q) &= \sum_{n=c}^{\infty}(n-c)\,p_n \\
&= \sum_{n=c}^{\infty}(n-c)\frac{\left(\frac{\lambda}{\mu}\right)^n}{c!\,c^{n-c}}\,p_0 \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{\lambda}{c\mu}\sum_{m=0}^{\infty} m\left(\frac{\lambda}{c\mu}\right)^{m-1} p_0 \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\,p_0\frac{\rho}{(1-\rho)^2} \\
&= \frac{\rho p_c}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}P(N \geq c) \\
&= \frac{\rho}{1-\rho}C(c,a). \qquad (3.6.7)
\end{aligned}
$$

This can also be found from the distribution of Q. Thus,

$$
\begin{aligned}
E(N) &= E(B) + E(Q) \\
&= c\rho + \rho\frac{p_c}{(1-\rho)^2} \\
&= c\rho + \frac{\rho C}{1-\rho}, \quad \text{where} \quad C = P\{N \geq c\} \equiv C(c,a). \qquad (3.6.8)
\end{aligned}
$$

$E\{N\}$ can be expressed in terms of Erlang's B formula (3.7.3) as

$$
E\{N\} = p_0\frac{a^c}{c!}\left[a\{(B(c,a)\}^{-1} - 1\} - c + \frac{1}{(1-\rho)^2}\right].
$$

Using Little's formula, we can find $E(W_Q)$, expected waiting time in the queue, and $E(W)$, expected waiting time in the system (sojourn or response time). We have

$$
E(W_Q) = \frac{E(Q)}{\lambda} = \frac{p_c}{c\mu(1-\rho)^2} = \frac{1}{c\mu(1-\rho)}P(N \geq c) \qquad (3.6.9)
$$

$$
\text{and} \quad E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{p_c}{c\mu(1-\rho)^2}. \qquad (3.6.10)
$$

Note: The relations (3.6.5) and (3.6.6) are fairly general and hold for any $G/G/c$ queue with $\rho = \lambda/c\mu < 1$. It can be seen simply as follows. We have

$$
L = \lambda W \quad \text{and}
$$

$$
L_Q = \lambda W_Q.
$$

Subtracting, we get $L - L_Q = \lambda(W - W_Q)$. The left-hand side gives the average number in the service channels or the average number of busy channels $E(B)$; $(W - W_Q)$ gives the average time spent in service and so equals $1/\mu$. Thus,

$$E(B) = \frac{\lambda}{\mu} = c\rho.$$

(For a direct proof, see Harris (1974).)

3.6.3 Waiting-time distributions

We can find the steady-state waiting-time distributions for the $M/M/c$ queue using the same arguments as were used for the $M/M/1$ system.

Let $w_q(x)$ and $w(x)$ be the PDFs of the waiting time W_q and W_s in the queue and in the system, respectively, of the test unit, and let $w_q^*(s)$ and $w^*(s)$ be their LTs. Further, let $w_q^*(s \mid n)$ and $w^*(s \mid n)$ be the LTs of the PDF of the conditional distributions of the respective waiting times given that the test unit finds, on arrival, n in the system. We obtain $w^*(s)$ by conditioning on the number of units the test unit finds on arrival. If the test unit finds, on arrival, $n < c$ units, he does not have to wait, and his waiting time in the system equals his service time—that is,

$$w^*(s \mid n) = \frac{\mu}{s + \mu}, \quad \text{for } n < c. \tag{3.6.11}$$

If he finds $n \geq c$ units in the system, he has to wait in the queue until the completion of service of $(n - c + 1)$ units; all the c service channels being busy, then the rate of service is $c\mu$. Taking into consideration his own service time, he thus has to wait in the system for completion of $(n - c + 1)$ services at the rate $c\mu$ and his own service at the rate μ, that is,

$$w^*(s \mid n) = \left(\frac{c\mu}{s + c\mu}\right)^{n-c+1} \left(\frac{\mu}{s + \mu}\right), \quad n \geq c. \tag{3.6.12}$$

Further, $a_n = p_n$ for this Poisson arrival process. Thus,

$$\begin{aligned} w^*(s) &= \sum_{n=0}^{c-1} w^*(s \mid n)\, p_n + \sum_{n=c}^{\infty} w^*(s \mid n)\, p_n \\ &= \frac{\mu}{s + \mu}\left[\sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} p_n \left(\frac{c\mu}{s + c\mu}\right)^{n-c+1}\right]. \end{aligned} \tag{3.6.13}$$

Putting in the values of p_n, we get

$$w^*(s) = \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \sum_{n=c}^{\infty}\frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}}\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right]p_0$$

$$= \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\left(\frac{c\mu}{s+c\mu}\right)\sum_{r=0}^{\infty}\left(\frac{\lambda}{c\mu}\right)^r\left(\frac{c\mu}{s+c\mu}\right)^r\right]p_0$$

$$= \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{(s+c\mu)}\frac{(s+c\mu)}{(s+c\mu-\lambda)}\right]p_0$$

$$= \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{s+c\mu-\lambda}\right]p_0. \tag{3.6.14}$$

We get $w(x)$ by inverting $w^*(s)$. From (3.6.14) we get

$$w^*(s) = \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\frac{\mu}{s+\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\left(\frac{\mu}{s+\mu}\right)\frac{c\mu}{s+c\mu-\lambda}$$

$$= \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\frac{\mu}{s+\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{c\mu^2}{(c-1)\mu-\lambda}\left[\frac{1}{s+\mu} - \frac{1}{s+c\mu-\lambda}\right].$$

Inverting the transform, we get

$$w(x) = \sum_{n+0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\mu e^{-\mu x} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{c\mu^2}{(c-1)\mu-\lambda}$$
$$\times\left[e^{-\mu x} - e^{-(1-\rho)c\mu x}\right].$$

For $c = 1$.

$$w(x) = \mu p_0 e^{-\mu x} + \frac{\left(\frac{\lambda}{\mu}\right)p_0\mu^2}{-\lambda}\left[e^{-\mu x} - e^{-(1-\rho)\mu x}\right]$$
$$= \mu(1-\rho)e^{-(1-\rho)\mu x}.$$

Since $w^*(s) = w_q^*(s)[\mu/(s+\mu)]$, we get from (3.6.13)

$$w_q^*(s) = \sum_{n=0}^{c-1}p_n + \sum_{n=c}^{\infty}p_n\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}$$

$$= \left[\sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{s+c\mu-\lambda}\right]p_0. \tag{3.6.15}$$

Inverting (3.6.15) we get the PDF

$$w_q(x) = \left(\sum_{n=0}^{c-1} p_n\right)\delta(x) + \sum_{n=c}^{\infty} p_n \frac{c\mu(c\mu x)^{n-c}e^{-c\mu x}}{(n-c)!}, \quad x \geq 0 \quad (3.6.16)$$

where δ is the Dirac delta (or unit impulse) function. Putting in the expressions for p_n and simplifying, we get

$$w_q(x) = \left(1 - \frac{p_c}{1-\rho}\right)\delta(x) + c\mu p_c e^{-c\mu(1-\rho)x}. \quad (3.6.17)$$

3.6.3.1 Complementary distribution function

We have, for $t \geq 0$,

$$\begin{aligned} Pr\{W_q > t\} &= \int_t^{\infty} w_q(x)dx \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{(c-1)!}\mu \frac{e^{-c(\mu-\lambda)t}}{c(\mu-\lambda)} p_0 \\ &= C\left(c, \frac{\lambda}{\mu}\right)e^{-(1-\rho)c\mu t} \\ &= \frac{\rho}{1-\rho} p_{c-1} e^{-(1-\rho)c\mu t}, \quad t \geq 0, \end{aligned} \quad (3.6.18)$$

where $C(c, \lambda/\mu) = Pr(W_q > 0)$ is the blocking probability, given by (3.6.4). Further,

$$\begin{aligned} Pr\{W_q > t \mid W_q > 0\} &= \frac{Pr\{W_q > t\}}{Pr\{W_q > 0\}} \\ &= e^{-(1-\rho)c\mu t}. \end{aligned} \quad (3.6.19)$$

The distribution of the conditional waiting time in the queue, given that the test unit has to wait, is exponential with mean

$$E\{W_q \mid W_q > 0\} = \frac{1}{(1-\rho)c\mu}. \quad (3.6.20)$$

Values of $E\{W_q\}$ and $E\{W_q \mid W_q > 0\}$ for certain values of c, λ, and μ are tabulated in Descloux (1962).

3.6.3.2 Mean and variance of waiting time

From (3.6.18), we get

$$E\{W_Q\} = [C(c,a)] \int_0^\infty e^{-(1-\rho)c\mu t}\, dt$$
$$= \frac{C(c,a)}{c\mu(1-\rho)} = \frac{C(c,a)}{c\mu - \lambda} \tag{3.6.21}$$

It can also be written as

$$E\{W_Q\} = \frac{p_c}{1-\rho} \cdot \frac{1}{c\mu - \lambda} = \frac{a^c}{c!c\mu(1-\rho)^2} p_0. \tag{3.6.21a}$$

Again, from (3.6.18)

$$E\{W_Q^2\} = 2C(c,a) \int_0^\infty te^{-(1-\rho)c\mu t}\, dt$$
$$= \frac{2C(c,a)}{(c\mu - \lambda)^2}$$

so that

$$\text{var}\{W_Q\} = \frac{2C(c,a)}{(c\mu - \lambda)^2} - \frac{\{C(c,a)\}^2}{(c\mu - \lambda)^2}$$
$$= \frac{C(c,a)[2 - C(c,a)]}{(c\mu - \lambda)^2}. \tag{3.6.22}$$

Now for the response time W_s, we get

$$E\{W_s\} = \frac{1}{\mu} + E\{W_Q\} = \frac{1}{\mu} + \frac{C(c,a)}{c\mu - \lambda}. \tag{3.6.23}$$

Similarly, we have

$$E(W_s^2) = \frac{d^2}{ds^2} w^*(s)|_{s=0}$$
$$= \frac{2}{\mu^2} + \left(\frac{2\left(\frac{\lambda}{\mu}\right)^c}{c!(c\mu)^2} \frac{p_0}{(1-\rho)^3} \right)$$
$$= \frac{2}{\mu^2} + \frac{2p_c}{(c\mu)^2(1-\rho)^3} = \frac{2}{\mu^2} + \frac{2C(c,a)}{[c\mu - \lambda]^2} \tag{3.6.24}$$

and the expression for variance of W_s can be written down.

Putting $c = 1$, we get the corresponding results for an $M/M/1$ queue. In particular, we have, for the $M/M/1$ queue,

$$E(W_s^2) = \frac{2}{\mu^2} + \frac{2(\frac{\lambda}{\mu})}{\mu^2(1-\rho)^2}.$$

Note: *Convexity properties* of performance measures (which have been found useful in the analysis of optimization problems (see Section 8.4)) have been receiving increasing attention. Grassman (1983) examines the convexity of $E(N)$ WRTρ. Harel and Zipkin (1987) show that $f_c(\rho) = 1/E(W_s)$ is strictly concave in ρ for fixed μ and $c \geq 2$, while $f_1(\rho) = \mu(1-\rho)$ is linear in ρ (for $c = 1$).

3.6.4 The output process

Here we consider the output process of an $M/M/c$ queueing system. The result is the same as that of an $M/M/1$ system, and the proof is similar.

Theorem 3.2. *In an M/M/c queueing system in steady state with rates of arrival and service λ and μ, respectively, the interdeparture times are independently and identically distributed as an exponential random variable with mean $1/\lambda$—that is, the output process is Poisson with parameter λ.*

Proof: Let $N(t)$ be the number in the system at time t and let $t_1', t_2', \ldots$ denote the successive departure instants, so that $L = t_{n+1}' - t_n'$ is the nth interdeparture interval. Let

$$F_k(t) = Pr\{N(t_n' + t) = k, \quad L > t\}, \quad t > 0, \quad k = 0, 1, 2, \ldots$$

We have, because of Poisson input,

$$\begin{aligned} F_k(0) = p_k &= \frac{(\frac{\lambda}{\mu})^k}{k!} p_0, \quad 0 \leq k \leq c, \\ &= \frac{(\frac{\lambda}{\mu})^k}{c!c^{k-c}} p_0, \quad k \geq c. \end{aligned} \tag{3.6.25}$$

For an infinitesimal interval of length dt

$$\begin{aligned} F_0(t + dt) &= F_0(t)[1 - \lambda\, dt] + o(dt) \quad \text{and} \\ F_k(t + dt) &= F_k(t)[1 - \lambda\, dt - j\mu\, dt] + F_{k-1}(t)\lambda\, dt + o(dt) \end{aligned} \tag{3.6.26}$$

where $j = k$ for $k < c$ and $j = c$ for $k \geq c$; the equations reduce to

$$\begin{aligned} F_0'(t) &= -\lambda F_0(t) \quad \text{and} \\ F_k'(t) &= \lambda F_{k-1}(t) - (\lambda + j\mu) F_k(t), \quad k = 1, 2, 3, \ldots \end{aligned} \tag{3.6.27}$$

Subject to the initial condition $F_k(0) = p_k$, the unique solution of (3.6.27) is given by

$$F_k(t) = p_k e^{-\lambda t}. \tag{3.6.28}$$

Again,

$$\begin{aligned} Pr\{N(t'_{n+1} + 0) = k,\ \ t \le t'_{n+1} - t'_n < t + dt\} \\ = F_{k+1}(t) \cdot Pr\{\text{one service completion in } (t, t + dt)\}. \end{aligned} \tag{3.6.29}$$

For $k + 1 \le c, (k + 1)(\le c)$ channels are busy and the rate of service of $(k + 1)\mu$, while for $k + 1 > c$, all the c channels are busy and the rate of service is $c\mu$. Thus, the RHS of (3.6.29) reduces to

$$\begin{aligned} F_{k+1}(t)(k + 1)\mu\ dt + o(dt), \quad k + 1 \le c, \quad \text{and} \\ F_{k+1}(t)c\mu\ dt + o(dt), \quad k + 1 > c. \end{aligned}$$

Now for $k + 1 \le c$.

$$\begin{aligned} F_{k+1}(t)[(k + 1)\mu\ dt] + o(dt) &= p_{k+1}e^{-\lambda t}[(k + 1)\mu\ dt] + o(dt) \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^{k+1}}{(k + 1)!} p_0 e^{-\lambda t}(k + 1)\mu\ dt + o(dt) \\ &= p_k \lambda e^{-\lambda t}\ dt + o(dt), \end{aligned}$$

and for $k + 1 > c$

$$\begin{aligned} F_{k+1}(t)c\mu\ dt + o(dt) &= p_{k+1}e^{-\lambda t}[c\mu\ dt] + o(dt) \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^{k+1}}{c!c^{k+1-c}} p_0 e^{-\lambda t} c\mu\ dt + o(dt) \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^{k}}{c!c^{k-c}} p_0 \lambda e^{-\lambda t}\ dt + o(dt) \\ &= p_k \lambda e^{-\lambda t}\ dt + o(dt). \end{aligned}$$

Thus,

$$\begin{aligned} Pr\{N(t'_{n+1} + 0) = k,\ \ t \le t'_{n+1} - t'_n < t + dt\} \\ = p_k \lambda e^{-\lambda t}\ dt + o(dt), \end{aligned} \tag{3.6.30}$$

which proves the independence of $N(t'_{n+1} + 0)$ and $t'_{n+1} - t'_n$, and it follows that $L = t'_{n+1} - t'_n$ has the density $\lambda e^{-\lambda t}$.

The proof of the independence of the interdeparture intervals is the same as that given for the $M/M/1$ queueing output process (given in Section 3.2.3). ■

Notes:

(1) The theorem is known as Burke's theorem. Burke (1968) also shows that the output process is independent of all other processes associated with the system and that of all the systems with FCFS service discipline, $M/M/c$ is the only system to possess this property.

(2) For an alternative proof of Burke's theorem and for generalizations, see Reich (1965).

3.7 The M/M/c/c System: Erlang Loss Model

Consider a c-server model with Poisson input and exponential service time such that when all the c-channels are busy an arrival leaves the system without waiting for service. This is called a (c-channel) *loss system* and was first investigated by Erlang.

This is a birth-and-death queueing model with

$$\begin{aligned} &\lambda_n = \lambda, \qquad \mu_n = n\mu, \quad n = 0, 1, 2, \ldots, c-1 \\ &\lambda_n = 0, \qquad \mu_n = c\mu, \quad n \geq c. \end{aligned} \tag{3.7.1}$$

Using (3.4.3) and (3.4.4) we get

$$\begin{aligned} p_n &= \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} p_0, \quad n = 1, \ldots, c \\ &= 0, \quad n > c \end{aligned}$$

and

$$p_0 = \left[\sum_{k=0}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} \right]^{-1}.$$

Thus,

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n / n!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k / k!}, \quad n = 0, 1, 2, \ldots, c. \tag{3.7.2}$$

The distribution of $\{p_n\}$ is truncated Poisson.

The above formula is known as Erlang's first formula. An arriving unit is lost to the system when he finds on arrival that all the channels are busy. The probability of this event is

$$p_c = \frac{\left(\frac{\lambda}{\mu}\right)^c / c!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k / k!}. \tag{3.7.3}$$

The above formula is known as *Erlang's loss* (or blocking, or overflow) *formula* or B-formula and is denoted by $B(c, \lambda/\mu)$.

Notes:

(1) While $a = \lambda/\mu$ is the *offered* load, $a' = a[1 - B(c, a)]$ is the *carried* load. The *overflow* rate of blocked (lost) customers is $a\,B(c, a)$: the sum of the carried load and the overflow rate equals the offered load a.

The *throughput*, defined as the rate at which customers (units) depart from the system after being (admitted and) served, is given by

$$a'\mu = \lambda[1 - B(c, a)];$$

this is the rate at which customers are accepted (admitted) for service (for equilibrium these two rates—acceptance rate and departure rate—must be equal).

A system in which customers have to leave when the space is full because of limited waiting space is called a *loss system*, whereas a system where all the arriving customers can wait (because of unlimited waiting space) is called a *delay* system. In a loss system, blocked customers (who arrive when the space is full) are said to be *cleared.*

(2) The formulas (3.7.2) and (3.7.3) hold irrespective of the form of the service time distribution—that is, these hold for an $M/G/c/c$ system (see Section 6.10.3.1 for a proof); and the service time distribution enters only through its mean μ. This is known as the *insensitivity* (or *robustness*) property.

Properties of Erlang loss formula (3.7.3) have been discussed by several researchers (e.g., Vaulot (1951), Fortet (1948), Jagerman (1974), Jagers and Van Doorn (1986), Takács (1969), Harel (1987), Berezner *et al.* (1998) among others; Medhi (2002) gives a survey). This aspect is also discussed in Section 6.10.3.1, where a proof of the formula (3.7.3) is outlined for the $M/G/c/c$ system.

(3) Newell (1984) discusses asymptotic approximations for Erlang's loss formula when the number of servers as well as the offered load are large.

(4) Messerli (1972) shows that Erlang's blocking probability [as given by B-formula (3.7.3)] is decreasing convex in c, the number of servers. This implies that the overflow or loss can be reduced by adding extra servers; however, the marginal decrease is itself decreasing. Yao (1986) considers a more general loss

system $G/M/m/m$ (with general interarrival time, ordered-entry, and heterogeneous servers) and examines convexity properties of the overflow.

In view of its importance, we again discuss Erlang loss function in Sections 3.7.1 and 3.7.2.

(5) Pacheco (1994a,b) discusses Loss Formulas for the model $M/M/c/c+r$, $r \geq 0$ (r denotes the number of excess space available when spaces before the c servers are all filled up). (See Problems and Complements 3.24.)

Example 3.1. Expected Number of Busy Channels
Let B be the RV denoting the number of busy channels. We have

$$\begin{aligned} E\{B\} &= \sum_{n=1}^{c} np_n = \sum_{n=1}^{c} \frac{n\left(\frac{\lambda}{\mu}\right)^n}{n!} p_0 \\ &= \left(\frac{\lambda}{\mu}\right) p_0 \sum_{n=1}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^{n-1}}{(n-1)!} = \left(\frac{\lambda}{\mu}\right) p_0 \left[\sum_{n=0}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} - \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} \right] \\ &= \frac{\lambda}{\mu}[1 - p_c] = \frac{\lambda}{\mu}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]. \end{aligned} \tag{3.7.4}$$

Thus, $E(B)$ equals the carried load a'.

If I is the RV denoting the number of idle channels, then

$$\begin{aligned} E\{I\} &= E\{c - B\} = c - E(B) \\ &= c - \frac{\lambda}{\mu}(1 - p_c) \\ &= c - \frac{\lambda}{\mu}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]. \end{aligned} \tag{3.7.5}$$

Example 3.2. Busy Probability (Kaufman, 1979)
Let X_i be the indicator variable for the ith randomly chosen channel: $X_i = 1$ or 0 according to whether the ith channel is busy or free. Let $P_c\{A\}$ denote the probability of an event A in an equilibrium $M/M/c/c$ loss system. Then

(i) $P_c\{X_1 = 1, \ldots, X_k = 1\} = \dfrac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c-k, \frac{\lambda}{\mu}\right)}, \quad 1 \leq k \leq c,$

(ii) $P_c\{X_1 = 1\} = \dfrac{\left(\frac{\lambda}{\mu}\right)\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right]}{c}$, and

(iii) $P_c\{X_{k+1} = 1 \mid X_1 = 1, X_2 = 1, \ldots, X_k = 1\} = P_{c-k}\{X_1 = 1\}.$

Proof (i): Conditioning on the number of busy channels, we get

$$P_c\{X_1 = 1, \ldots, X_k = 1\} = \sum_{j=k}^{c} Pr\{X_1 = 1, \ldots, X_k = 1 \mid n = j\} p_j$$

$$= \sum_{j=k}^{c} \frac{\binom{j}{k}}{\binom{c}{k}} \; p_0 \frac{\left(\frac{\lambda}{\mu}\right)^j}{j!} = \frac{(c-k)!}{c!} p_0 \sum_{j=k}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^j}{(j-k)!}$$

$$= \frac{(c-k)!}{c!}\left(\frac{\lambda}{\mu}\right)^{k-c} \sum_{j=k}^{c}\left[\left(\frac{\lambda}{\mu}\right)^c p_0\right]\frac{\left(\frac{\lambda}{\mu}\right)^{j-k}}{(j-k)!}$$

$$= \left[\frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} p_0\right]\left[\frac{(c-k)!}{\left(\frac{\lambda}{\mu}\right)^{c-k}} \sum_{r=0}^{c-k} \frac{\left(\frac{\lambda}{\mu}\right)^r}{r!}\right]$$

$$= \frac{Pr\{c \text{ channels busy in } M/M/c/c\}}{Pr\{c-k \text{ channels busy in } M/M/c-k/c-k\}}$$

$$= \frac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c-k, \frac{\lambda}{\mu}\right)}. \tag{3.7.6}$$

■

Proof (ii):

$$P_c\{X_1 = 1\} = \sum_{j=1}^{c} P_c\{X_1 = 1 \mid B = j\}\, Pr\{B = j\}$$

($B \equiv$ number of busy channels)

$$= \sum_{j=1}^{c} \frac{j}{c} p_j = \frac{1}{c} \sum_{j=1}^{c} j p_j$$

$$= \frac{1}{c} E(B)$$

$$= \frac{\lambda/\mu}{c}\left[1 - B\left(c, \frac{\lambda}{\mu}\right)\right] \quad \text{by Eq. (3.7.4).} \tag{3.7.7}$$

■

Proof (iii): Again,

$$P_c\{X_{k+1} = 1 \mid X_1 = 1, X_2 = 1, \ldots, X_k = 1\}$$

$$= \frac{P_c\{X_1 = 1, X_2 = 1, \ldots, X_k = 1, X_{k+1} = 1\}}{P_c\{X_1 = 1, X_2 = 1, \ldots, X_k = 1\}}$$

$$= \frac{B\left(c, \frac{\lambda}{\mu}\right)}{B\left(c-k-1, \frac{\lambda}{\mu}\right)} \frac{B\left(c-k, \frac{\lambda}{\mu}\right)}{B\left(c, \frac{\lambda}{\mu}\right)} = \frac{B\left(c-k, \frac{\lambda}{\mu}\right)}{B\left(c-k-1, \frac{\lambda}{\mu}\right)}$$

$$= \frac{\left[\left(\frac{\lambda}{\mu}\right)^{c-k}(c-k)!\right]\left[\sum_{r=0}^{c-k}\left(\frac{\lambda}{\mu}\right)^r/r!\right]^{-1}}{\left[\left(\frac{\lambda}{\mu}\right)^{c-k-1}(c-k-1)!\right]\left[\sum_{r=0}^{c-k-1}\left(\frac{\lambda}{\mu}\right)^r/r!\right]^{-1}}$$

$$= \frac{\frac{\lambda}{\mu}}{c-k}\left[1 - \frac{\left(\frac{\lambda}{\mu}\right)^{c-k}/(c-k)!}{\sum_{r=0}^{c-k}\left(\frac{\lambda}{\mu}\right)^r/r!}\right]$$

$$= \frac{\frac{\lambda}{\mu}}{c-k}\left[1 - B\left(c-k, \frac{\lambda}{\mu}\right)\right]$$

$$= P_{c-k}\{X_1 = 1\}. \tag{3.7.8}$$

Now (iii) implies that if $(k+1)$ channels in an $M/M/c/c$ system in equilibrium are randomly chosen without replacement and the first k channels are busy, then the conditional probability that the $(k+1)$st channel is busy equals the *a priori* probability that one randomly chosen channel in an $M/M/c-k/(c-k)$ system is busy. ■

Note: Kaufman shows that the results of Examples 3.1 and 3.2 also hold for the loss systems with Poisson input (with rate λ) and general service-time distribution (with mean rate μ).

3.7.1 Erlang loss (blocking) formula: Recursive algorithm

In practical problems, occasions arise to compute $B(c, a)$ for values of c and a.

A numerically convenient formula for calculating $B(c, a)$ is via

$$[B(c,a)]^{-1} = \sum_{j=0}^{\infty} c^{(j)} a^{-j}, \tag{3.7.9}$$

where

$$c^{(j)} = \begin{cases} 1, & j = 0 \\ c(c-1)\ldots(c-j+1), & j \geq 1. \end{cases}$$

It is useful when $a > c$. Now

$$a^{-j} = a\int_0^{\infty} e^{-ay}\frac{y^j}{j!}\,dy, \quad a > 0, \quad j \geq 0.$$

Thus, (3.7.9) admits of Fortet's integral representation

$$[B(c,a)]^{-1} = a\int_0^{\infty} e^{-ay}(1+y)^c\,dy$$

$$= \int_0^{\infty} e^{-y}\left(1 + \frac{y}{a}\right)^c dy. \tag{3.7.10}$$

Equation (3.7.10) permits evaluation of loss function for non-integral number c of servers in practical traffic-related problems. The sensitivity of the formula can be extended to complex c and a by means of the interpolatory analytic function

$$[B(x,a)]^{-1} = \int_0^\infty e^{-y}\left(1+\frac{y}{a}\right)^x dy, \quad Re(a) > 0. \tag{3.7.11}$$

The corresponding extension of (3.7.9) is the asymptotic expansion, as $a \to \infty$,

$$[B(x,a)]^{-1} \sim \sum_{j=0}^{\infty} x^{(j)} a^{-j}, \quad |\arg a| < \pi. \tag{3.7.12}$$

Further, integration of (3.7.10) or (3.7.11) by parts gives the useful *recurrence relation*

$$[B(x+1,a)]^{-1} = \frac{x+1}{a}[B(x,a)]^{-1} + 1. \tag{3.7.13}$$

For integral $x = n$, $B(n,a)$ can be conveniently computed from (3.7.13) from the initial value $B(0,a) = 1$.

The recursive algorithm (3.7.13) can also be used for non-integral x by using the integral representation (3.7.10).

3.7.2 Relation between Erlang's *B* and *C* formulas

We recall that while $B(c,a)$ (for real a, integral c) holds for any value of $\rho = a/c = \lambda/c\mu$ and also for general service time distribution, $C(c,a)$ holds only for $\rho < 1 (a < c)$ and for $M/M/c$ system only.

For every integer $c > a$, it can be easily verified that $C(c,a)$ can be expressed in terms of $B(c,a)$ as follows:

$$C(c,a) = \frac{B(c,a)}{1-\rho\{1-B(c,a)\}}. \tag{3.7.14}$$

Since $B(c,a)$ for integral c can be conveniently computed using the recursive algorithm (3.7.13) starting with $B(0,a) = 1$, one can compute $C(c,a)$ with the help of (3.7.14).

Since the denominator of the RHS of (3.7.14) is a proper fraction, we have

$$C(c,a) > B(c,a). \tag{3.7.15}$$

Graphs of $C(c,a)$ as a function of a for different integral values of c are given in Cooper (1981).

Remarks: The transient behavior of Loss Model $M/M/c/c$ has been discussed, among others, by Knessl (1990), Abate and Whitt (1998), and Fricker *et al.* (1999).

See Problems and Complements 3.25.

3.8 Model with Finite Input Source

3.8.1 Steady-state distribution: $M/M/c//m$ $(m > c)$. Engset delay model

Our assumption so far has been that the source of population from which the arrivals occur is infinite. Now we shall examine the case of arrivals from a source with finite population—say, of size m. A unit may be in the system or outside the system. The system consists of a fixed number—say, of c parallel servers—where $c < m$. A unit entering the system starts receiving service from one of the parallel servers if there is any free server available or joins the queue, if there is none. The service time distribution of each of the c servers is IID exponential with parameter μ. The total service rate when n servers are busy is $n\mu$, when $n \leq c$ and $c\mu$, when $n \geq c$. If at any instant there are n in the system, (either receiving service or $n-c(\geq 0)$ in the queue while c are receiving service), then there are $(m - n)$ outside the system from which arrivals to the system occur, the average arrival rate being $\lambda(m - n)$; the distribution of interarrival time is IID exponential with parameter λ. We have state-dependent arrival and service rates. To fix our ideas, we can consider the following example. Suppose that there are m machines and c repairpersons (servers) to repair them, when required. A machine is in the system when it is in a failed (or nonworking) state requiring repair facility of one of the c servers. When all the repair-persons are busy, the machine joins the queue. The time required to repair a machine is IID exponential with parameter μ. A machine in working order is outside the system, the time for breakdown (or life time) of a machine being exponential with parameter λ. The system may be denoted by $M/M/c//m$ (Kleinrock, 1975).

The situation can be appropriately described by a birth-death model with state-dependent rates (of arrivals into the system having n units therein and $(m - n)$ outside)

$$\begin{aligned} \lambda_n &= (m-n)\lambda, \quad n = 0, 1, \ldots, m-1 \\ &= 0, \quad n \geq m \end{aligned}$$

and (of departures from the system having n units therein)

$$\begin{aligned} \mu_n &= n\mu, \quad n = 1, 2, \ldots, c-1 \\ &= c\mu, \quad n \geq c. \end{aligned} \tag{3.8.1}$$

Assume that the system (the whole system) is in steady state. Denote

$$p_n = Pr\{\text{the number in the system is } n\}.$$

Using (3.4.3), we have, for $n = 0, 1, \ldots, c-1$,

$$\begin{aligned} p_n &= \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} \, p_0 = \prod_{i=0}^{n-1} \frac{(m-i)\lambda}{(i+1)\mu} \, p_0 \\ &= \frac{m(m-1)(m-2)\ldots(m-n+1)}{n!} \left(\frac{\lambda}{\mu}\right)^n p_0 \\ &= \binom{m}{n} \left(\frac{\lambda}{\mu}\right)^n p_0 \end{aligned} \tag{3.8.2}$$

and for $n = c, c+1, \ldots, m$,

$$\begin{aligned} p_n &= \prod_{i=0}^{n-1} \frac{(m-i)\lambda}{(i+1)\mu} \, p_0 \\ &= \frac{m!}{(m-n)!} \cdot \frac{1}{c!c^{n-c}} \cdot \left(\frac{\lambda}{\mu}\right)^n p_0. \end{aligned} \tag{3.8.3}$$

Using the normalizing condition

$$\sum_{n=0}^{m} p_n = 1,$$

we get

$$p_0 = \left[\sum_{n=0}^{c-1} \binom{m}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{m} \frac{m!}{(m-n)!} \cdot \frac{1}{c!c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1} \tag{3.8.4}$$

Notes:

(1) In the context of the machine repairperson problem,

$$p_n = Pr\{\text{number of machines under or awaiting repair is } n\}$$
$$q_n = Pr\{\text{number of machines in working order is } n\} = p_{m-n}.$$

(2) When $\rho < 1$, then taking limit as $m \to \infty$, we get the results of the $M/M/c$ queue.

(3) The distribution of the number k of busy servers in an Engset delay model $M/M/c//m$ $(m > c)$ is given by

$$\begin{aligned} p_k &= Pr\{\text{number of busy servers} = k\}, \quad k = 0, 1, 2, \ldots, c \\ &= \binom{m}{k} \left(\frac{\lambda}{\mu}\right)^k p_0 \\ p_0 &= \left[\sum_{i=0}^{c} \binom{m}{i} \left(\frac{\lambda}{\mu}\right)^i \right]^{-1}. \end{aligned} \tag{3.8.5}$$

This is known as Engset distribution.

(4) Bunday and Scarton (1980) show that the results [(3.8.2), (3.8.3), (3.8.4)] hold for any system having finite input source and having exponential service time (with parameter μ) and general independent identical interarrival times or lifetimes (with parameter λ)—that is, for a $G/M/c$ model with finite input source of size m.

Example 3.3. An Application of the $M/M/N//N$ Model in Electronics (Koenigsberg, 1980)

Semiconductor noise as a queueing problem was first formulated by Bell. Koenigsberg discusses cyclic queue models of semiconductor noise. One model considered by him is a finite-input-source model with N customers (electrons) and N servers (impurity levels), where N is large. Each server can serve one customer at a time, and each customer in the queue can be served by one vacant or free server. The situation is similar to having N machines subject to breakdown and repairs by N repairpeople.

Let

x = rate of electron withdrawal from the conduction band (rate of breakdown of working machines), which depends on the number of vacant impurity levels,

y = rate of electron excitation to the conduction band (rate of repair of broken-down machines), which depends on the number of busy impurity levels, and

$p_n = Pr\{n \text{ electrons are in the conduction band}\}$.

Then putting $m = c = N, \lambda = x, \mu = y$ in (3.8.2) and (3.8.4), we have

$$p_n = \binom{N}{n}\left(\frac{x}{y}\right)^n p_0, \quad n = 0, 1 \ldots, N-1, N,$$

$$p_0 = \left[\sum_{n=0}^{N}\binom{N}{n}\left(\frac{x}{y}\right)^n\right]^{-1} = \left(1+\frac{x}{y}\right)^{-N},$$

$$\text{and} \quad L = \sum_{n=0}^{N} np_n = \frac{Nx}{y}\left(1+\frac{x}{y}\right)^{-1} = \frac{Nx}{x+y}.$$

Notes:

Some other models of electron excitation are considered by Koenigsberg. The second model is equivalent to that of the machine-interference problem, in which machines subject to breakdown can be repaired in the "transition zone" by one of the N repairpeople; however, at any time $(N - m)$ repairpeople are engaged in other duties and are not available for repairs. The other duties arise in accordance with a Poisson process of rate—say, z—and completions take place in accordance with an independent Poisson process of rate—say, s.

The third model is equivalent to one in which the repairperson moves the machine from the point of breakdown to a repair facility where it is placed in service (excitation). Then the repairperson returns to the base, perhaps carrying out other duties en route.

3.8.1.1 Waiting-time distribution for an M/M/c//m model ($m > c$)

Consider the exponential model described in the preceding section with c servers and input from a finite source of m units ($m > c$). Wong (1979) obtained the waiting-time distribution for such a model.

Let a_n be the probability that an arrival finds n units in the system (and $m-n$ outside): in case of a machine-interference problem, a_n is the probability that n machines are not in working order. Here the input is quasi-random, being from a finite source. We can find a_n from (3.3.6) on replacing λ by $(m-n)\lambda$; then

$$a_n = \frac{(m-n)\,p_n}{\sum_{k=0}^{m}(m-k)\,p_k} = \frac{(m-n)\,p_n}{m-L}, \tag{3.8.6}$$

where $L = \sum kp_k$ is the average number of machines in the system that are not in working order and $m-L$ is the average number of machines in working order.

Define $w^*(s)$ to be the LST of the waiting time in the system (response time). The LST $w^*(s)$ can be obtained by conditioning on the number of units n that an arrival finds. Define $w^*(s \mid n)$ to be the LST of the conditional distribution of the waiting time in the system (response time) given that the test unit finds n in the system. Let $w_q^*(s)$ and $w_q^*(s \mid n)$ be the corresponding quantities for waiting time in the queue. We have

$$w^*(s) = \sum_{n=0}^{m-1} w^*(s \mid n)a_n.$$

Now

$$\begin{aligned} w^*(s \mid n) &= \frac{\mu}{s+\mu} \quad \text{for } n < c \\ &= \left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\left(\frac{\mu}{s+\mu}\right) \quad \text{for } c \le n \le m-1 \end{aligned} \tag{3.8.7}$$

and

$$w^*(s) = \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1} a_n + \sum_{n=c}^{m-1} a_n\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right], \tag{3.8.8}$$

where a_n is given by (3.8.6). It follows that

$$w_q^*(s) = \sum_{n=0}^{c-1} a_n + \sum_{n=c}^{m-1} a_n \left(\frac{c\mu}{s + c\mu} \right)^{n-c+1}. \tag{3.8.9}$$

Inverting the transform, we get

$$w_q(x) = \left(\sum_{n=0}^{c-1} a_n \right) \delta(x) + \sum_{n=c}^{m-1} a_n \frac{(c\mu)(c\mu x)^{n-c} e^{-c\mu x}}{(n-c)!}. \tag{3.8.10}$$

Inversion of $w^*(s)$ is done next. One can find the moments of W_s from (3.8.8).

3.8.1.2 Inversion of $w^*(s)$

The inversion can be carried in two ways: (i) as the inversion of the convolution and (ii) as the inversion of terms obtained by resolution of the expression into partial fractions. The method given by Jordan (1950, Section 13c, p. 38) may be used for resolution into partial fractions. We consider the first method next.

Consider the inversion of

$$\frac{\mu}{s + \mu} \left(\frac{c\mu}{s + c\mu} \right)^{n-c+1}.$$

The inverse transform of the first factor

$$f_1^*(s) = \frac{\mu}{\mu + s}$$

is

$$f_1(x) = \mu e^{-\mu x} \tag{3.8.11}$$

and that of the second factor

$$f_2^*(s) = \left(\frac{c\mu}{s + c\mu} \right)^{n-c+1}$$

is

$$f_2(x) = (c\mu)^{n-c+1} \frac{x^{n-c} e^{-c\mu x}}{(n-c)!}. \tag{3.8.12}$$

Using these and noting that the inverse of $f_1^*(s) f_2^*(s)$ is the convolution of $f_1(x)$ and $f_2(x)$, we find that the inverse LT $L^{-1}[f_1^*(s) f_2^*(s)]$ of $[f_1^*(s) f_2^*(s)]$

is given by

$$L^{-1}[f_1^*(s)\, f_1^*(s)] = \mu(c\mu)^{n-c+1}\left[\int_0^x e^{-\mu(x-t)}\frac{t^{n-c}e^{-c\mu t}}{(n-c)!}dt\right]$$

$$= \frac{\mu(c\mu)^{n-c+1}}{(n-c)!}e^{-\mu x}\int_0^x t^{n-c}e^{-\mu(c-1)t}\,dt.$$

Using the representation of the incomplete gamma function

$$\int_0^x \frac{\alpha^k t^{k-1}e^{-\alpha t}}{\Gamma(k)} = 1 - \sum_{r=0}^{k-1} e^{-\alpha x}\frac{(\alpha x)^r}{r!} \tag{3.8.13}$$

we get

$$L^{-1}\left[\frac{\mu}{s+\mu}\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right]$$

$$= \frac{\mu(c\mu)^{n-c+1}e^{-\mu x}}{(n-c)![\mu(c-1)]^{n-c+1}}\left[1 - \sum_{r=0}^{n-c} e^{-\mu(c-1)x}\frac{[\mu(c-1)x]^r}{r!}\right]$$

$$= \mu\left(\frac{c}{c-1}\right)^{n-c+1} e^{-\mu x}\left[1 - e^{\mu x}\sum_{r=0}^{n-c} e^{-c\mu x}\frac{[\mu(c-1)x]^r}{r!}\right]. \tag{3.8.14}$$

Thus, we find from (3.8.8), (3.8.11), and (3.8.14) that the inverse Laplace transform $w(x)$ of $w^*(s)$ given by (3.8.8) can be written as

$$w(x) = \mu e^{-\mu x}\sum_{n=0}^{c-1} a_n + \sum_{n=c}^{m-1} a_n\left[\left(\frac{c}{c-1}\right)^{n-c+1}\mu e^{-\mu x}\right.$$

$$\left. - \mu\left(\frac{c}{c-1}\right)^{n-c+1}\sum_{r=0}^{n-c} e^{-\mu x}\frac{[\mu(c-1)x]^r}{r!}\right]. \tag{3.8.15}$$

The expression for the distribution function $F(x) = \int_0^x w(t)dt$ can be written down.

Note: The model has been used to describe various important situations arising out of *machine-interference problems* and to find solutions of design problems such as determination of the optimal number of repairpeople for a given set of machines, the maximum length of time during which all m machines are working, and so on.

3.8.2 Engset loss model $M/M/c//m/(m > c)$

Here we suppose that the finite population size m is greater than the number of servers c and there is no place for waiting for an arrival (a failed unit) when all the c servers are occupied. Then an arrival (a newly failed unit) who does

not find a free server is lost to the system (consisting of only those with the servers); the system size n takes values $0, 1, 2, \ldots, c$.

We then have (from 3.8.5)

$$p_n = \frac{\binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n}{\sum_{i=0}^{c}\binom{m}{i}\left(\frac{\lambda}{\mu}\right)^i}, \quad n = 0, 1, \ldots, c, \tag{3.8.16}$$

which is *Engset distribution.*

The loss probability is given by

$$p_c = \frac{\binom{m}{c}a^c}{\sum_{i=0}^{c}\binom{m}{i}a^i}, \quad a = \lambda/\mu. \tag{3.8.17}$$

3.8.2.1 Blocking probabilities

There are two types of blocking: those due either to time congestion or to call congestion. *Time congestion* is defined as the proportion of time that the system is blocked (that is, no new arrival can be admitted into the system). *Call congestion* is defined as the proportion of calls that are blocked.

For Poisson arrivals, it follows from PASTA that the steady-state distribution of the number in the system as found by an arrival is the same as the steady-state distribution of the number in the system at an arbitrary point of time. Thus, time congestion and call congestion (and their probabilities) are the same. Thus, for $M/M/c/c$ (also for $M/G/c/c$) system, the blocking probability (for both time and call congestion) is given by Erlang Loss Formula

$$B(c, a) = \frac{a^c/c!}{G(c)}, \quad \text{where}$$

$$a = \lambda/\mu \quad \text{is the offered load and}$$

$$G(c) = \sum_{j=0}^{c}\frac{a^j}{j!} \quad \text{is the normalization constant.}$$

Consider now a Poisson arrival stream from a finite population, m—that is, an Engset system $M/M/c//m$. If $m \leq c$, there can be no blocking. If $m > c$, the time congestion is given by

$$p_c = T(m, c, a) = \binom{m}{c}a^c/G(m, c), \quad \text{where} \tag{3.8.18}$$

$$G(m, c) = \sum_{i=0}^{c}\binom{m}{i}a^i \quad \text{is the normalization constant.}$$

Here, PASTA does not hold. We have to find the call congestion. The expected number of arrivals when there are n busy servers is

$$\{(m-n)\lambda\}T(m,n,a), \quad n=0,1,\ldots,c.$$

The expected number of arrivals per unit time is

$$\sum_{n=0}^{c}\{(m-n)\lambda\}T(m,n,a)$$

Thus, call congestion T_c is given by

$$\begin{aligned} T_c(m,c,a) &= \frac{\{(m-c)\lambda\}T(m,c,a)}{\sum_{n=0}^{c}\{(m-n)\lambda\}T(m,n,a)} \\ &= \binom{m-1}{c}a^c/G(m-1,c) \\ &= T(m-1,c,a) \end{aligned} \tag{3.8.19}$$

so that call congestion for an Engset model with finite population of size m equals the time congestion for the corresponding Engset model with population size $(m-1)$.

3.8.2.2 Recursive formula for Engset loss formula

We have obtained a recursive algorithm for the Erlang Loss formula in Section 3.7.1. Here we consider a recursive formula for the blocking probability for the Engset model.

We have

$$G(m,c)=\sum_{i=0}^{c}\binom{m}{i}a^i=\binom{m}{c}a^c+G(m,c-1)$$

so that

$$\frac{G(m,c)}{\binom{m}{c}a^c}=1+\frac{G(m,c-1)}{\binom{m}{c}a^c}.$$

That is,

$$[T(m,c,a)]^{-1}=1+\frac{c}{a(m-c+1)}[T(m,c-1,a)]^{-1}. \tag{3.8.20}$$

With initial condition $T(m,0,a)=1$, the above yields a recursive algorithm for computing $T(m,c,a)$ for fixed values of a,m, and intermediate values of c. (See also Kobayashi and Mark (1997).)

3.8.3 The model $M/M/c//m$ $(m \leq c)$

Here the number n in the system is the same as the number of busy servers. We have

$$\lambda_i = \lambda(m-i), \quad 0 \leq k \leq m$$
$$= 0, \quad \text{otherwise}$$
$$p_n = \prod_{i=0}^{n-1} \frac{\lambda(m-i)}{(i+1)\mu} p_0, \quad n = 0, 1, \ldots, m$$
$$= \binom{m}{n} \left(\frac{\lambda}{\mu}\right)^n p_o.$$

Now

$$\sum_{n=0}^{m} p_n = 1 \quad \text{gives}$$
$$p_0 = \left[\sum_{n=0}^{m} \binom{m}{n} \left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$$
$$= \left[\left(1 + \frac{\lambda}{\mu}\right)^m\right]^{-1} = \frac{1}{\left(1 + \frac{\lambda}{\mu}\right)^m}. \tag{3.8.21}$$

Thus,

$$p_n = \frac{\binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n}{\left(1 + \frac{\lambda}{\mu}\right)^m}, \quad n = 0, 1, 2, \ldots, m$$
$$= \binom{m}{n} \left(\frac{\lambda}{\mu}\right)^n \left[\left(1 + \frac{\lambda}{\mu}\right)^{-1}\right]^m. \tag{3.8.22}$$

This can also be written as

$$p_n = \binom{m}{n} \left[\frac{\mu}{\lambda+\mu}\right]^{m-n} \left[\left(\frac{\lambda}{\mu}\right)\left(\frac{\mu}{\lambda+\mu}\right)\right]^n, \quad m-n = 0, 1, 2, \ldots, m$$
$$= \binom{m}{n} \left(\frac{\mu}{\lambda+\mu}\right)^{m-n} \left(\frac{\lambda}{\lambda+\mu}\right)^n, \quad m-n = 0, 1, \ldots, m$$
$$= \binom{m}{n} (1-p)^{m-n} (p)^n, \tag{3.8.22a}$$

where $p = \lambda/(\lambda+\mu)$.

Thus the distribution is binomial with

$$p = \frac{\lambda}{\lambda + \mu} = \frac{1/\mu}{1/\lambda + 1/\mu}$$
$$= P\{\text{a unit is in service at a service station}\} \qquad (3.8.23)$$

(which follows from a property of alternating renewal process—see Section 1.8).

We shall have the same result, even if the finite population size is m and the number of servers c is infinite or the number of servers c is equal to the finite population size m.

That is, the result holds, for the models

$$M/M/\infty//m \quad \text{and} \quad M/M/m//m.$$

Remarks: As mentioned earlier, the Erlang model is insensitive to the service (holding) time distribution and depends on it only through its mean $1/\mu$—that is, insensitivity applies to the $M/G/c/c$ model. Insensitivity with respect to service time distribution applies also to Engset models. Further insensitivity also applies to interarrival time for the Engset loss model—that is, to the model $GI/G/c/c/m$ $(m > c)$. (See Kobayashi and Mark (1997).)

3.9 Transient Behavior

3.9.1 Introduction

Most of the known results in queueing theory pertain to steady state. This is because the equations involved become considerably simplified in the limit when time t from the initialization becomes very large. Analytically tractable results can then be obtained. Nevertheless, steady-state conditions do not hold well in many applied situations as the time horizon of operation terminates and remains finite. For example, the repairperson at a service facility leaves after a fixed duration of time; so also is the case of a bankteller. Steady-state analysis and performance measures obtained thereof do not then make much sense. Transient behavior is more meaningful under such circumstances: an analysis that deals with a system's operating behavior for a fixed, finite amount of time and takes into account the initial conditions is more relevant.

Transient results are, however, more difficult to obtain, and not many such results are found in the literature. Morse (1955) studied the $M/M/1$ queue and obtained the transient state probabilities of the number in the system at time t. The problem was studied, and a complete solution was obtained in the 1950s by several researchers using different methods (as indicated in the next section). Transient state behavior of $M/M/s$ queue was studied by Saaty (1960) and Jackson and Henderson (1966). Transient characteristics of $M/M/\infty$ have recently been studied, among others by Gullemin and Simonian (1995).

Note: In the 1980s, there had been a fresh interest in the transient behavior of $M/M/1$ queue as indicated by such papers as Abate and Whitt (1987, 1988), Hubbard *et al.* (1986), Parthasarathy (1987), Syski (1988), Pegden and Rosenshine (1982), Sharma and Gupta (1982), and Sharma (1997).

Similar interest has also been evinced in case of $M/M/c$ queue. See, for example, Van Doorn (1981, Chapter 6), Kelton and Law (1985), and Parthasarathy and Sharafali (1989).

While most authors consider the continuous-time framework, Kelton and Law (1985) carry out their analysis of transient behavior in discrete time through indexing of customer number. (See Problems and Complements 3.20.)

We examine below transient state behavior through continuous-time framework.

3.9.1.1 Transient state distribution for the M/M/1 model

This single-server model envisages Poisson input and exponential service time with FCFS queue discipline. The arrivals occur in accordance with a Poisson process with parameter (or intensity)—say, λ—that is, the probability of one arrival in an infinitesimal interval of length h is $\lambda h + o(h)$, while that of more than one arrival is $o(h)$. The distribution of service time is exponential with parameter—say, μ—that is, the probability of one service completion in an interval of infinitesimal length h is $\mu h + o(h)$, and that of more than one service completion is $o(h)$. The model corresponds to that of a birth-death process with rates $\lambda_n = \lambda, n = 0, 1, 2, \ldots$ and $\mu_n = \mu, n = 1, 2, \ldots$ Denote

$$p_n(t) = Pr\{N(t) = n\}, \quad n \geq 0,$$

where $N(t)$ is the number in the system at time t. Using (1.4.5) and (1.4.6) of Chapter 1, it can be easily seen that $p_n(t)$ satisfies the differential-difference equations given below.

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \tag{3.9.1}$$

$$p_n'(t) = -(\lambda + \mu) p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), \quad n \geq 1. \tag{3.9.2}$$

These equations in transient state have been solved by using a number of techniques: by the spectral method (Ledermann and Reuter, 1954), by the generating function method (Bailey, 1954), by the combinatorial method (Champernowne, 1956), and by the difference equation solution technique (Conolly, 1958; Feller, 1966). See Saaty (1961); see also Conolly and Langaris (1993), Laguesdron *et al.* (1993) and Problems 3.13, 3.14, and 3.15 for some recent alternative approaches. We consider below the difference-equation technique and the method of generating function.

3.9.2 Difference-equation technique

Knowledge of the initial state is required for obtaining transient state distribution. For simplicity, suppose that the number of units present at $t = 0$ is 0—that is, at time $t = 0$ the system is empty. This implies that $p_0(0) = 1$ and $p_n(0) = 0, n \neq 0$. Let $\bar{p}_n(s)$ be the LT of $p_n(t)$. Then taking LT of (3.9.1) and (3.9.2) and using

$$L\{p_n'(t)\} = sL\{p_n(t)\} - p_n(0),$$

we get

$$(s + \lambda)\bar{p}_0(s) = 1 + \mu\bar{p}_1(s) \quad \text{and} \tag{3.9.3}$$

$$(s + \lambda + \mu)\bar{p}_n(s) = \lambda\bar{p}_{n-1}(s) + \mu\bar{p}_{n+1}(s), \quad n \geq 1. \tag{3.9.4}$$

The equation (3.9.4) is a difference equation of order two, having for its characteristic equation

$$\mu z^2 - (s + \lambda + \mu)z + \lambda = 0. \tag{3.9.5}$$

Let z_1, z_2 be the roots of (3.9.5)

$$z_i \equiv z_i(s) = \frac{(s + \lambda + \mu) \pm \sqrt{\{(s + \lambda + \mu)^2 - 4\lambda\mu\}}}{2\mu}$$

with $i = 1$ (with +ve sign before the radical), $i = 2$ (with −ve sign before the radical). We have

$$z_1 + z_2 = (s + \lambda + \mu)/\mu, \qquad z_1 z_2 = \lambda/\mu, \quad |z_1| > |z_2|.$$

Further, $|z_1| > 1$ and $|z_2| < 1$ as can be seen by applying the following.

Rouché's Theorem. *If $f(z)$ and $g(z)$ are functions analytic inside and on a closed contour C, and if $|g(z)| < |f(z)|$ on C, then $f(z)$ and $f(z) + g(z)$ have the same number of zeros inside C.*

Here we take C as the unit circle $|z| = 1$, $f(z) = (s + \lambda + \mu)z, g(z) = \lambda + \mu z^2$, $\text{Re}(s) > 0$. It follows that

$$\begin{aligned} |f(z)| &= |(s + \lambda + \mu)z| = |s + \lambda + \mu| \\ &\geq |\lambda + \mu| = \lambda + \mu \\ &\geq |\lambda + \mu z^2| = |g(z)|. \end{aligned}$$

As $f(z)$ has only one zero inside $|z| = 1$, the equation

$$f(z) + g(z) = \mu z^2 - (s + \lambda + \mu)z + \lambda = 0$$

will also have only one root inside C.

Now the root z_2 being of smaller modulus must be such that $|z_2| < 1$. The solution of the difference equation (3.9.4) can be written as

$$\bar{p}_n(s) = Az_1^n + Bz_2^n, \quad n \geq 1. \tag{3.9.6}$$

Taking LT of

$$\sum_{n=0}^{\infty} p_n(t) = 1,$$

we get

$$\sum \bar{p}_n(s) = 1/s.$$

Thus, $\sum \bar{p}_n(s) = \sum(Az_1^n + Bz_2^n)$ converges and, since $|z_1| > 1$, A must be identically equal to 0, so that

$$\bar{p}_n(s) = Bz_2^n, \quad n \geq 1.$$

We can choose B to yield the correct value of $\bar{p}_0(s)$ such that (3.9.4) is satisfied for $n = 1$; this gives $\bar{p}_0(s) = B$. Thus, we have

$$\bar{p}_n(s) = \bar{p}_0(s)z_2^n, \quad n \geq 0. \tag{3.9.7}$$

Using $\sum \bar{p}_n(s) = 1/s$, we get

$$\bar{p}_0(s) = (1 - z_2)/s$$

and

$$\bar{p}_n(s) = (1 - z_2)z_2^n/s, \quad n \geq 0. \tag{3.9.7a}$$

3.9.2.1 Steady-state distribution

When it exists, the steady-state probability p_n can be obtained by taking the limit of $p_n(t)$ as $t \to \infty$. We have

$$\begin{aligned} \lim_{s\to 0} z_2 &= \lim_{s\to 0} \frac{(s + \lambda + \mu) - \sqrt{\{(s + \lambda + \mu)^2 - 4\lambda\mu\}}}{2\mu} \\ &= \begin{cases} \frac{(\lambda+\mu)-(\mu-\lambda)}{2\mu} = \rho, & \text{when } \lambda < \mu \\ \frac{(\lambda+\mu)-(\lambda-\mu)}{2\mu} = 1, & \text{when } \lambda \geq \mu. \end{cases} \end{aligned} \tag{3.9.8}$$

By applying the initial value theorem of Laplace transforms, we get

$$\begin{aligned} p_n &= \lim_{t\to\infty} p_n(t) = \lim_{s\to 0} s\,\bar{p}_n(s) \\ &= \lim_{s\to 0}(1 - z_2)z_2^n \\ &= \begin{cases} (1-\rho)\rho^n & \rho < 1 \\ 0, & \rho \geq 1 \end{cases} \qquad n = 0, 1, 2, \ldots \end{aligned} \tag{3.9.9}$$

The interpretation of the result $p_n = 0, \rho \geq 1$ is that when the traffic intensity is greater than or equal to 1, then the probability that the system contains a finite number of units n is zero, as should be intuitively clear (because of the increasing queue length). Note that when $\rho = 1$, the states are persistent null with infinite recurrence time.

3.9.2.2 Transient-state distribution

Assume that $p_0(0) = 1$. We have from (3.9.7a)

$$\begin{aligned} \bar{p}_n(s) &= \frac{(1 - z_2)z_2^n}{s} \\ &= \frac{(1 - z_2)z_2^n}{\{\mu(z_1 - 1)(1 - z_2)\}}, \quad \text{since } -s = \mu(1 - z_1)(1 - z_2) \\ &= \left(\frac{z_2^{n+1}}{\lambda}\right)\left[1 + \sum_{r=0}^{\infty}\left(\frac{1}{z_1}\right)^{r+1}\right], \quad \text{since } |1/z_1| < 1 \\ &= \left(\frac{1}{\lambda}\right)\left[z_2^{n+1} + \sum_{r=0}^{\infty} z_2^{n+1}\left(\frac{\mu}{\lambda}\right)^{r+1} z_2^{r+1}\right] \\ &= \left(\frac{1}{\lambda}\right)\left[z_2^{n+1} + \left(\frac{\lambda}{\mu}\right)^{n+1} \sum_{k=n+2}^{\infty}\left(\frac{\mu}{\lambda}\right)^k z_2^k\right]. \end{aligned} \tag{3.9.10}$$

We need the inverse Laplace transform of z_2^n to find $p_n(t)$. We have

$$\begin{aligned} z_2 &= \frac{(s + \lambda + \mu) - \sqrt{\{(s + \lambda + \mu)^2 - 4\lambda\mu\}}}{(2\mu)} \\ &= \frac{4\lambda\mu}{2\mu[(s + \lambda + \mu) + \sqrt{\{(s + \lambda + \mu)^2 - 4\lambda\mu\}}} \end{aligned}$$

Writing

$$\frac{4\lambda\mu}{2\mu} = 2\sqrt{\lambda\mu}\sqrt{\frac{\lambda}{\mu}}$$

we get

$$z_2^n = \left(\frac{\lambda}{\mu}\right)^{n/2}\left[\frac{2\sqrt{\lambda\mu}}{(s+\lambda+\mu)+\sqrt{\{(s+\lambda+\mu)^2-4\lambda\mu\}}}\right]^n.$$

From the Table of Laplace transform, we note that

$$\text{LT of}\left\{\frac{n}{t}I_n(at)\right\} \quad \text{is} \quad \left\{\frac{a}{s+\sqrt{(s^2-a^2)}}\right\}^n.$$

Using the translation property of LT, we get that LT of

$$\left[\left(\frac{\lambda}{\mu}\right)^{n/2} e^{-(\lambda+\mu)t}\left\{\frac{n}{t}I_n(2t\sqrt{\lambda\mu})\right\}\right] \text{ is } z_2^n. \tag{3.9.11}$$

Hence, from (3.9.10) and (3.9.11) we get

$$p_n(t) = \frac{e^{-(\lambda+\mu)t}}{\lambda}\left[\left(\frac{\lambda}{\mu}\right)^{(n+1)/2}\left(\frac{n+1}{t}\right)I_{n+1}(at) + \left(\frac{\lambda}{\mu}\right)^{n+1}\sum_{k=n+2}^{\infty}\left(\frac{\mu}{\lambda}\right)^{k+2}\left(\frac{k}{t}\right)I_k(at)\right], \tag{3.9.12}$$

where $2\sqrt{\lambda\mu} = a$.

Using the property

$$\frac{2m}{z}I_m(z) = I_{m-1}(z) - I_{m+1}(z),$$

we get

$$\frac{2k}{at}I_k(at) = I_{k-1}(at) - I_{k+1}(at).$$

Substituting in (3.9.12), we get

$$p_n(t) = e^{-(\lambda+\mu)t}\frac{\sqrt{\lambda\mu}}{\lambda}\left[\left(\frac{\lambda}{\mu}\right)^{(n+1)/2}\{I_n(at) - I_{n+2}(at)\} + \left(\frac{\lambda}{\mu}\right)^n\left\{\sum_{k=n+2}^{\infty}\left(\frac{\mu}{\lambda}\right)^{(k-2)/2}(I_{k-1}(at) - I_{k+1}(at))\right\}\right]$$

$$= e^{-(\lambda+\mu)t}\left[\left(\frac{\lambda}{\mu}\right)^{n/2} I_n(at) - \left(\frac{\lambda}{\mu}\right)^{n/2} I_{n+2}(at)\right.$$
$$+ \left(\frac{\lambda}{\mu}\right)^{n}\left\{\left(\frac{\mu}{\lambda}\right)^{(n+1)/2} \{I_{n+1}(at) - I_{n+3}(at)\}\right.$$
$$\left.\left. + \left(\frac{\mu}{\lambda}\right)^{(n+2)/2} \{I_{n+2}(at) - I_{n+4}(at)\}\right\} + \cdots\right]$$
$$= e^{-(\lambda+\mu)t}\left[\left(\frac{\mu}{\lambda}\right)^{-n/2} I_n(at) + \left(\frac{\mu}{\lambda}\right)^{(-n+1)/2} I_{n+1}(at)\right.$$
$$+ \left(\frac{\lambda}{\mu}\right)^{n}\left(1 - \frac{\lambda}{\mu}\right)\left\{\left(\frac{\mu}{\lambda}\right)^{(n+2)/2} I_{n+2}(at)\right.$$
$$\left.\left. + \left(\frac{\mu}{\lambda}\right)^{(n+3)/2} I_{n+3}(at) + \cdots\right\}\right].$$

Finally,

$$p_n(t) = e^{-(\lambda+\mu)t}\left[\rho^{n/2} I_n(at) + \rho^{(n-1)/2} I_{n+1}(at)\right.$$
$$\left. + \{(1-\rho)\rho^n\}\left\{\sum_{k=n+2}^{\infty} \rho^{-k/2} I_k(at)\right\}\right], \quad n \geq 0. \tag{3.9.13}$$

So far we assumed that $p_0(0) = 1$. The general case $p_i(0) = 1$, $p_j(0) = 0$, $j \neq i$ can be treated similarly; when $p_i(0) = 1$, we shall get, for all $n \geq i$, $n < i$.

$$p_n(t) = e^{-(\lambda+\mu)t}\left[\rho^{(n-i)/2} I_{n-i}(at) + \rho^{(n-i-1)/2} I_{n+i+1}(at)\right.$$
$$\left. + (1-\rho)\rho^n \sum_{k=n+i+2}^{\infty} \rho^{-k/2} I_k(at)\right] \quad n \geq 0. \tag{3.9.14}$$

See also Problems and Complements 3.13–3.15.

3.9.2.3 Derivation of steady-state distribution

Assume that $\rho < 1$. Using the limiting property as $t \to \infty$,

$$\lim I_r(at) \to \frac{\exp(at)}{\sqrt{(2\pi at)}} \quad \text{(independent of } r\text{)},$$

it can be shown that, as $t \to \infty$, the first two terms of both the expressions on the RHS of (3.9.13) and (3.9.14) tend to 0 and the third term tends to unity. Thus, when $\rho < 1$,

$$p_n = \lim_{t\to\infty} p_n(t) = (1-\rho)\rho^n, \quad n \geq 0.$$

3.9.3 Method of generating function

We now consider the popular method introduced by Bailey (1954).

Denote

$$P(z,t) = \sum_{n=0}^{\infty} p_n(t) z^n, \quad |z| < 1$$

$$\tilde{P}(z,s) = \text{LT of } P(z,t) = \sum_{n=0}^{\infty} \tilde{p}_n(s) z^n.$$

Multiplying (3.9.2) by z^n, $n = 1, 2, \ldots$ and adding with (3.9.1), we get

$$\sum_{n=0}^{\infty} p_n'(t) z^n = -(\lambda + \mu) \sum_{n=0}^{\infty} p_n(t) z^n + \mu p_0(t) + \lambda \sum_{n=1}^{\infty} p_{n-1}(t) z^n + \mu \sum_{n=0}^{\infty} p_{n+1}(t) z^n$$

or

$$\frac{\partial}{\partial t} P(z,t) = -(\lambda + \mu) P(z,t) + \mu p_0(t) + \lambda z P(z,t) + \left(\frac{\mu}{z}\right)\{P(z,t) - p_0(t)\}$$

or

$$z \frac{\partial}{\partial t} P(z,t) = \{\lambda z^2 - (\lambda + \mu) z + \mu\} P(z,t) - \mu(1 - z) p_0(t). \quad (3.9.15)$$

Taking LT and noting that

$$L\left\{\frac{\partial}{\partial t} P(z,t)\right\} = s \tilde{P}(z,s) - P(z,0)$$

we get

$$z[s \tilde{P}(z,s) - P(z,0)] = \{\lambda z^2 - (\lambda + \mu) z + \mu\} \tilde{P}(z,s) - \mu(1 - z) \tilde{p}_0(s)$$

whence

$$\tilde{P}(z,s) = \frac{\mu(1 - z) \tilde{p}_0(s) - z P(z,0)}{\lambda z^2 - (s + \lambda + \mu) z + \mu}. \quad (3.9.16)$$

The relation involves $P(z,0)$ (which can be found from the initial condition) as well as $\tilde{p}_0(s)$. To find $\tilde{p}_0(s)$, we note that the denominator of the RHS of

(3.9.16) has two roots—say, ξ_1 and ξ_2,

$$\xi_i = \frac{(s+\lambda+\mu) \mp \sqrt{\{(s+\lambda+\mu)^2 - 4\lambda\mu\}}}{2\lambda}, \quad i = 1,2.$$

Now the roots of $\lambda z^2 - (s+\lambda+\mu)z + \mu = 0$ are the reciprocals of the roots of the equation $\mu z^2 - (s+\lambda+\mu)z + \lambda = 0$. Thus, $\xi_1 = 1/z_1, \xi_2 = 1/z_2, |\xi_1| < 1$, $|\xi_2| > 1$.

As $\bar{P}(z,s)$ converges in the region $|z| \leq 1$, the zero (in the unit disc) of the numerator and denominator of $\bar{P}(z,s)$ must coincide. Thus, $z = \xi_1$ must also be a root of the numerator, so that

$$\mu(1-\xi_1)\bar{p}_0(s) = \xi_1 P(\xi_1, 0)$$

or

$$\bar{p}_0(s) = \frac{\xi_1 P(\xi_1, 0)}{\mu(1-\xi_1)}, \tag{3.9.17}$$

and finally, putting the value of $\bar{p}_0(s)$ in (3.9.16), we get

$$\bar{P}(z,s) = \frac{(1-z)\xi_1 P(\xi_1,0) - z(1-\xi_1)P(z,0)}{\lambda(z-\xi_1)(z-\xi_2)(1-\xi_1)}. \tag{3.9.18}$$

$P(z,0)$ can be found from the initial condition. Suppose that $p_i(0) = 1$, then $P(z,0) = z^i$ and

$$\bar{P}(z,s) = \frac{(1-z)\xi_1^{i+1} - z^{i+1}(1-\xi_1)}{\lambda(z-\xi_1)(z-\xi_2)(1-\xi_1)}. \tag{3.9.19}$$

3.9.3.1 Particular case

When $i = 0$, $p_0(0) = 1$, $p_n(0) = 0$, $n \neq 0$, and $P(z,0) = 1$, so also $P(\xi_1, 0) = 1$. Then

$$\begin{aligned}
\bar{P}(z,s) &= \frac{(1-z)\xi_1 - z(1-\xi_1)}{\lambda(z-\xi_1)(z-\xi_2)(1-\xi_1)} \\
&= \frac{z_1 z_2}{\lambda(z_1-1)} \sum_{n=0}^{\infty} z_2^n z^n \\
&= \frac{(1-z_2)}{s} \sum_{n=0}^{\infty} z_2^n z^n, \quad \text{since} \quad -s = \mu(z_1-1)(z_2-1).
\end{aligned}$$

Hence,

$$\begin{aligned}
\bar{p}_n(s) &= \text{coefficient of } z^n \text{ in } \bar{P}(z,s) \\
&= \frac{(1-z_2)z_2^n}{s}, \quad n = 0,1,2,\ldots \text{ (as in 3.9.7a).}
\end{aligned}$$

Remarks:

(1) Abate and Whitt (1988) show how the transform analysis can be continued to obtain a better description of the transient behavior; they discuss time-dependent performance measures (1989).

(2) The approach through application of Rouché's theorem, which can be used for some other models as well, could be quite complicated in practice. Neuts (1979) formulates an alternative approach for queues solvable without Rouché's theorem.

3.9.4 Busy-period analysis

We define a busy period as the interval of time from the instant a unit arrives at an empty system and its service begins, to the instant when the server becomes *free* for the *first time.* A busy period is an RV, being the first passage time from state 1 to state 0. Denote

$$\begin{aligned} T &= \text{length of the busy period} \\ b(t) &= \text{PDF of } T \\ N^*(t) &= \text{number present at time } t \text{ during a busy period} \\ &\quad \{N^*(t), t \geq 0\} \text{ is a zero-avoiding state process} \\ q_n(t) &= Pr\{N^*(t) = n\}, \quad n = 0, 1, 2, \ldots \\ \tilde{q}_n(s) &= \text{LT of } q_n(t). \end{aligned}$$

We have $q_1(0) = 1, q_n(0) = 0, n \neq 1$, for $n \geq 2, q_n(t)$ will satisfy the same differential equations as $p_n(t)$—that is, Eq. (3.9.2) will hold good also for $q_n(t), n = 2, 3, \ldots$ Thus,

$$q_n'(t) = -(\lambda + \mu)q_n(t) + \lambda q_{n-1}(t) + \mu q_{n+1}(t), \quad n \geq 2. \tag{3.9.20}$$

As the term $q_0(t)$ will not occur, the equation corresponding to $n = 1$ will be

$$q_1'(t) = -(\lambda + \mu)q_1(t) + \mu q_2(t). \tag{3.9.21}$$

Taking LT of (3.9.20)

$$\mu\tilde{q}_{n+1}(s) - (s + \lambda + \mu)\tilde{q}_n(s) + \lambda\tilde{q}_{n-1}(s) = 0, \quad n \geq 2. \tag{3.9.22}$$

This is a difference equation of order 2, having the same characteristic equation (3.9.5). Thus,

$$\tilde{q}_n(s) = Az_1^n + Bz_2^n, \quad n \geq 2.$$

Since $\sum \bar{q}_n(s)$ converges and $|z_1| > 1$, $A \equiv 0$. We can choose the constant B such that (3.9.22) is satisfied for $n = 2$. Thus, $\bar{q}_1(s) = Bz_2$ and

$$\bar{q}_n(s) = \bar{q}_1(s) z_2^{n-1}, \quad n \geq 1. \tag{3.9.23}$$

The LT of (3.9.21) yields

$$s\bar{q}_1(s) = 1 - (\lambda + \mu)\bar{q}_1(s) + \mu z_2 \bar{q}_1(s)$$

or

$$[(s + \lambda + \mu) - \mu z_2]\bar{q}_1(s) = 1.$$

Thus

$$\begin{aligned} \bar{q}_1(s) &= \frac{1}{(s + \lambda + \mu) - \mu z_2} \\ &= \frac{z_2}{\lambda}, \quad \text{(since } z_2 \text{ is a root of (3.9.5))} \end{aligned}$$

so that

$$\bar{q}_n(s) = \bar{q}_1(s) z_2^{n-1} = \frac{z_2^n}{\lambda}.$$

Inversion of the LT yields

$$q_n(t) = \left(\frac{\lambda}{\mu}\right)^{n/2} \frac{n}{\lambda t} e^{-(\lambda+\mu)t} I_n(2t\sqrt{\lambda\mu}), \quad n = 1, 2, \ldots \tag{3.9.24}$$

Conditioning on the number of units present at instant t, all of which complete their service in $(t, t + dt)$, we have

$$\begin{aligned} b(t)dt &= Pr\{t \leq T < t + dt\} \\ &= \sum_{j=1}^{\infty} Pr\{t \leq T < t + dt \mid N^*(t) = j\} \times Pr\{N^*(t) = j\} \\ &= Pr\{t \leq T < t + dt \mid N^*(t) = 1\}\, Pr\{N^*(t) = 1\} \\ &\quad + \sum_{j=2}^{\infty} Pr\{t \leq T < t + dt \mid N^*(t) = j\}\, Pr\{N^*(t) = j\} \end{aligned}$$

The first term implies that there is only one unit (at the instant t) whose service is completed between $(t, t + dt)$, the probability of this event being $\mu dt + o(dt)$. The second term implies service completion of two or more units in $(t, t + dt)$

and the probability of this event is $o(dt)$. Thus, taking limit as $dt \to 0$,

$$\begin{aligned} b(t) &= [\mu q_1(t)] \\ &= \frac{1}{t}\rho^{-1/2}e^{-(\lambda+\mu)t}I_1(2t\sqrt{\lambda\mu}). \end{aligned} \tag{3.9.25}$$

The LST of T is given by

$$\begin{aligned} b^*(s) &= L\{b(t)\} = \mu\tilde{q}_1(s) \\ &= \mu\frac{z_2}{\lambda} = \frac{z_2}{\rho}. \end{aligned} \tag{3.9.26}$$

Now

$$\begin{aligned} \lim_{s\to 0} z_2 &= \rho \quad \text{if } \rho < 1 \\ &= 1 \quad \text{if } \rho \geq 1 \end{aligned}$$

so that

$$\begin{aligned} b^*(0) &= 1 \quad \text{if } \rho < 1 \\ &= \frac{1}{\rho} \quad \text{if } \rho \geq 1. \end{aligned}$$

This shows that there is non-zero probability that the busy period, when $\rho > 1$, is infinitely large, as should be intuitively clear.

3.9.4.1 Moments of the busy period

Writing

$$K = \sqrt{\{(s+\lambda+\mu)^2 - 4\lambda\mu\}},$$

we have

$$\frac{d}{ds}z_2 = [1 - 2(s+\lambda+\mu)/2K]/(2\mu) = -z_2/K$$

so that

$$\begin{aligned} E(T) &= -\frac{d}{ds}b^*(s)|_{s=0} \\ &= \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}. \end{aligned}$$

Again

$$\begin{aligned} \frac{d^2}{ds^2}z_2 &= \frac{z_2}{K^2} + \frac{z_2}{K^3}(s+\lambda+\mu) \\ &= \frac{2\lambda}{K^3} \end{aligned}$$

so that

$$E(T^2) = \frac{d^2}{ds^2} b^*(s)|_{s=0} = \frac{2\lambda}{\rho} \frac{1}{(\mu - \lambda)^3}$$
$$= \frac{2}{\mu^2(1-\rho)^3}.$$

We have

$$\text{var}(T) = \frac{1+\rho}{\mu^2(1-\rho)^3}.$$

Remarks:

(1) It was the celebrated French mathematician Emile Borel (1871–1956) who introduced the concept of busy period. He obtained the joint distribution of the busy period T and the number N served during the busy period for an $M/D/1$ model.

(2) The LST of the busy period can also be obtained from a certain functional equation that it satisfies (see Section 6.4.2).

(3) *Idle period* is the interval I from the instant the server becomes free to the instant of the next arrival (when the server resumes service). This interval is the residual interarrival time. Since the interarrival time is exponential, the idle period I is also exponential, with the same parameter λ as the interarrival time. Thus, $E(I) = 1/\lambda$ for a model with Poisson input.

(4) The expected duration of the busy period can also be obtained from a result of renewal theory. (See Section 1.8, relation (1.8.1).) The busy period T and the idle period I form an alternating renewal process. Thus,

$$\frac{E(T)}{E(I)} = \frac{1-p_0}{p_0}.$$

Since $p_0 = 1 - \rho$ for a single-server system, we get

$$E(T) = \frac{\rho}{1-\rho}\left(\frac{1}{\lambda}\right) = \frac{1}{\mu(1-\rho)}, \quad (\rho = a = \lambda/\mu).$$

The result holds for a system with general service time distribution.

(5) We have examined the busy period initiated by a single unit. The busy period initiated by r units (with $r \geq 1$ units in the system at the commencement of the busy period) is given by the sum of r IID random variables $T_i (\equiv T$, the busy period initiated by a single unit). Unless otherwise stated, we shall take $r = 1$.

(6) Let $E(N)$ denote the average number of customers served during a busy period T. For every $E(N)$ arrivals during a busy period exactly one arrival (the first customer during a busy period) will find the system empty. Hence, the probability a_0 that an arrival finds the system empty is given by

$$a_0 = \frac{1}{E(N)}.$$

As PASTA holds, $a_0 = p_0 = 1 - \rho$ so that $E(N) = 1/(1 - \rho)$.

We can get the result intuitively as follows: since the server remains continuously busy serving during a busy period, we have

$$\begin{aligned} E(N) &= E(T) \times \{\text{rate of service}\} \\ &= \frac{1}{\mu(1-\rho)} \times \mu \\ &= \frac{1}{1-\rho}. \end{aligned} \tag{3.9.27}$$

3.9.4.2 Number served during a busy period of an M/M/1 queue

Let N be the number served during a busy period that starts with one customer. We find the distribution of N.

Consider the system-size process observed at each arrival and departure epoch as a one-dimensional random walk with a reflecting barrier at the origin. Let X_n be the system size at the nth arrival departure epoch. The transition probabilities are given by

$$\begin{aligned} p_{i,i+1} &= Pr\{X_{n+1} = i+1 \mid X_n = i\} \\ &= Pr\{\text{an arrival occurs before a departure occurs}\} \\ &= \frac{\lambda}{\lambda+\mu} = \frac{\rho}{1+\rho} = p \quad \text{(say)}, \quad \text{and} \end{aligned} \tag{3.9.28}$$

$$\begin{aligned} p_{i,i-1} &= Pr\{X_{n+1} = i-1 \mid X_n = i\} \\ &= Pr\{\text{a departure occurs before an arrival occurs}\} \\ &= \frac{\mu}{\lambda+\mu} = \frac{1}{1+\rho} = 1 - p. \end{aligned} \tag{3.9.29}$$

The event that $\{N = n\}$ is equivalent to the event that the *first* return to the origin through the positive axis of the random walk occurs at epoch $2n$. Now $Pr\{N = n\} = Pr\{\text{the first return to origin occurs at epoch } 2n\}$ is given by

$$\varphi_{1,2n-1} = \frac{1}{2n-1}\binom{2n-1}{\frac{2n-1+1}{2}} p^n (1-p)^{n-1} \tag{3.9.30}$$

(Feller, 1968, vol. I, Th. 4, p. 90). That is,

$$Pr\{N = n\} = \frac{1}{2n-1}\binom{2n-1}{n}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}$$
$$= \frac{1}{n}\binom{2n-2}{n-1}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}, \quad n = 1, 2, 3, \ldots \tag{3.9.31}$$

The PGF of N is given by

$$P(s) = \sum_{n=1}^{\infty} Pr(N = n)s^n = \frac{2s}{1+\rho+[(1+\rho)^2 - 4\rho s]}. \tag{3.9.32}$$

We have

$$E(N) = \frac{1}{(1-\rho)} \tag{3.9.33}$$

$$\text{and} \quad \text{var}(N) = \frac{\rho(1+\rho)}{(1-\rho)^3}. \tag{3.9.34}$$

When the busy period starts with m customers, then

$$Pr\{N = n\} = \frac{m}{n}\binom{2n-m-1}{n-1}\frac{\rho^{n-m}}{(1+\rho)^{2n-m}}, \quad n = m, m+1, \ldots. \tag{3.9.35}$$

Its PGF is

$$P(s) = \sum_{n=m}^{\infty} Pr(N = n)s^n$$
$$= \frac{(2s)^m}{[1+\rho+\{(1+\rho)^2 - 4\rho s\}]^m} \tag{3.9.36}$$

$$\text{with} \quad E(N) = \frac{m}{(1-\rho)} \tag{3.9.37}$$

$$\text{and} \quad \text{var}(N) = \frac{m\rho(1+\rho)}{(1-\rho)^3}. \tag{3.9.38}$$

Haight (1961) has described the distribution of N which is "analogous to the Borel-Tanner distribution" relating to the number served during the busy period of an $M/D/1$ queue.

Example 3.3. First passage time distribution

For an $M/M/1$ queue in transient state, let τ be the random variable denoting the time taken for the system to *fall* from the initial state a to another state $b(< a)$ for the first time. Let $\{X(t)\}$ be the process denoting such states of the

system, given $X(0) = a$, and let

$$q_n(t) = P\{X(t) = n \mid X(0) = a\}. \tag{3.9.39}$$

Using similar arguments, it can be easily shown that the state probabilities satisfy the following equations:

$$\begin{aligned} q_n'(t) &= -(\lambda + \mu)q_n(t) + \lambda q_{n-1}(t) + \mu q_{n+1}(t), \quad n \geq (b+2) \\ q_{b+1}'(t) &= -(\lambda + \mu)q_{b+1}(t) + \mu q_{b+2}(t) \\ q_b'(t) &= \mu q_{b+1}(t). \end{aligned} \tag{3.9.40}$$

The PDF $f(t)$ of the first passage time distribution from state a to state b is given by $f(t) = \mu q_{b+1}(t) = q_b'(t)$. Putting $a = 1$ and $b = 0$, we can get back to the initial busy-period distribution.

3.9.5 Waiting-time process: Virtual waiting time

Let $W(t)$ be the time required to serve all the units present in the system at the instant t, given that $W(0) = 0$. If $N(t)$ is the number present at instant t, $(N(0) = 0)$, then

$$W(t) = \begin{cases} 0 & \text{if } N(t) = 0 \\ v_1' + v_2 + \cdots + v_{N(t)} & \text{if } N(t) > 0, \end{cases}$$

where v_1' is the residual service time of the unit being served at the instant t, and $v_2, \ldots v_{N(t)}$ are the service times of the units waiting at the instant t. $\{W(t), t > 0\}$, which is known as *virtual waiting time*, is a Markov process (with continuous-state space). Given $W(0) = 0$, then proceeding as in the case of waiting time in the system in steady state, it can be seen that its probability element $f(x,t)dx = P\{x \leq W(t) < x + dx\}, 0 < x < \infty, 0 < t < \infty$, is

$$f(x,t)dx = \sum_{n=0}^{\infty} \mu \frac{(\mu x)^n}{\Gamma(n+1)} e^{-\mu x}\, dx\, p_n(t).$$

Its LT can be put in a closed form. (See Prabhu (1965).)

Example 3.4. Transient solution of the $M/M/1/1$ model

Here $\lambda_0 = \lambda, \mu_0 = 0$, and $\lambda_1 = 0, \mu_1 = \mu$; if $N(t)$ denotes the number in the system at time t, then $Pr\{N(t) = n\} = p_n(t) = 0$ for all $n > 1$—that is, we are concerned with only $p_0(t)$ and $p_1(t)$ such that $p_0(t) + p_1(t) = 1$. The differential-difference equations of the model then become

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t) \\ p_1'(t) &= -\mu p_1(t) + \lambda p_0(t). \end{aligned}$$

Writing $p_1(t) = 1 - p_0(t)$ in the first equation, we get $p_0'(t) + (\lambda + \mu)p_0(t) = \mu$. The solution of this first-order linear differential equation with constant coefficients is given by

$$p_0(t) = Ce^{-(\lambda+\mu)t} + \frac{\mu}{(\lambda+\mu)},$$

where C is constant. Given the initial distribution $p_i(0) = Pr\{N(0) = i\}$, we get

$$p_0(t) = p_0(0)e^{-(\lambda+\mu)t} + \frac{\mu}{\lambda+\mu}\left\{1 - e^{-(\lambda+\mu)t}\right\}.$$

Similarly,

$$p_1(t) = p_1(0)e^{-(\lambda+\mu)t} + \frac{\lambda}{\lambda+\mu}\left\{1 - e^{-(\lambda+\mu)t}\right\}.$$

The steady-state solutions are

$$p_0 = \lim_{t\to\infty} p_0(t) = \frac{\mu}{(\lambda+\mu)}$$

$$p_1 = \lim_{t\to\infty} p_1(t) = \frac{\lambda}{(\lambda+\mu)},$$

irrespective of whether the value of $\rho = \lambda/\mu < 1$ or not.

Assume that the initial distribution is identical with the steady-state distribution so that

$$p_0(0) = p_0 = \frac{\mu}{(\lambda+\mu)} \quad \text{and} \quad p_1(0) = p_1 = \frac{\lambda}{(\lambda+\mu)}.$$

Then we find that for all $t > 0$,

$$p_0(t) = \frac{\mu}{(\lambda+\mu)} = p_0 \quad \text{and} \quad p_1(t) = \frac{\lambda}{(\lambda+\mu)} = p_1.$$

That is, if the process is in equilibrium (steady state) initially, then it will be always (for all $t > 0$) in steady state. This is true for any other ergodic system.

Notes: In case of the $M/M/c/c$ model, if the system is in equilibrium initially—that is, $Pr\{N(0) = n\} = p_n(0) = p_n, 0 < n < c$, where p_n are steady-state probabilities (given by relation(3.7.2)), then $\{N(t), t > 0\}$ becomes a stationary process for which $N(t)$ has the same distribution for all $t > 0$. That is,

$$p_n(t) = Pr\{N(t) = n\} = p_n$$

(given by (3.7.2)) for all $t > 0$. (See Takács (1969).)

3.10 Transient-State Distribution of the $M/M/c$ Model

3.10.1 Solution of the differential-difference equations

Consider a c-server queueing system with Poisson input (with parameter λ) and exponential service time (with parameter μ) for each of the c-servers. The model corresponds to that of a birth-death process with rates

$$\begin{aligned}\lambda_n &= \lambda, \quad && n = 0, 1, 2, \ldots,\\ \mu_n &= n\mu, && n = 1, 2, \ldots, c-1,\\ &= c\mu, && n = c, c+1, c+2, \ldots.\end{aligned}$$

Let $N(t)$ be the number in the system at time t and let

$$p_n(t) = Pr\{N(t) = n\}.$$

Then p_n's satisfy the differential-difference equations

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \tag{3.10.1}$$

$$p_n'(t) = -(\lambda + n\mu)\, p_n(t) + \lambda p_{n-1}(t) + (n+1)\mu p_{n+1}(t), \quad 1 \le n \le c-1, \tag{3.10.2}$$

$$p_n'(t) = -(\lambda + c\mu)\, p_n(t) + \lambda p_{n-1}(t) + c\mu p_{n+1}(t), \quad n \ge c. \tag{3.10.3}$$

These equations in transient state have been solved by Saaty (1960) and Jackson and Henderson (1966). We consider here the difference-equation technique of Jackson and Henderson. Denote the initial condition by

$$p_n(0) = \delta_{in} (i \text{ being the number at time } 0).$$

Consider that $i = c$ at time 0; time is reckoned from the instant when all the servers become busy with none in the queue. Let $p_n^*(s)$ denote the LT of $p_n(t)$. Taking the LT of (3.10.1) through (3.10.3), we get

$$(\lambda + s)\, p_0^*(s) = \mu p_1^*(s), \tag{3.10.4}$$

$$(\lambda + s + n\mu)\, p_n^*(s) = \lambda p_{n-1}^*(s) + (n+1)\mu p_{n+1}^*(s), \quad 1 \le n \le c-1, \tag{3.10.5}$$

$$(\lambda + s + c\mu)\, p_n^*(s) - \delta_{cn} = \lambda p_{n-1}^*(s) + c\mu p_{n+1}^*(s), \quad n \ge c. \tag{3.10.6}$$

The second difference equation is one with variable coefficients, and as such it has to be solved by using special techniques as follows. Assume that solutions exist—that is, the equations are consistent. The solutions $p_n^*(s), 0 \le n \le c-1$ of the first two equations can be obtained independently of $p_n^*(s), n \ge c$.

We shall use some interesting properties of generating functions. Let us write $p_n^*(s) = f(n)$; then (3.10.5) can be written as

$$(n+1)\mu f(n+1) = (\lambda + s + n\mu) f(n) - \lambda f(n-1). \tag{3.10.7}$$

Let

$$G[f(n)] = \sum_{n=0}^{\infty} f(n)t^n = F(t) \tag{3.10.8}$$

be the generating function of $\{f(n)\}$. Multiplying both sides of (3.10.7) by t^n for $n = 1, 2, \ldots$ and adding, we get

$$\mu \sum_{n=1}^{\infty}(n+1) f(n+1)t^n = (\lambda + s)\sum_{n=1}^{\infty} f(n)t^n + \mu \sum_{n=1}^{\infty} nf(n)t^n - \lambda t \sum_{n=1}^{\infty} f(n-1)t^{n-1}. \tag{3.10.9}$$

We have

$$\sum_{n=1}^{\infty} nf(n)t^n = t\sum_{n=1}^{\infty} nf(n)t^{n-1} = tF'(t)$$

$$\sum_{n=1}^{\infty}(n+1) f(n+1)t^n = \sum_{m=1}^{\infty} mf(m)t^{m-1} - f(1) \quad (\text{putting } n+1 = m)$$

$$= F'(t) - f(1).$$

Thus, from (3.10.9), we have

$$\mu[F'(t) - f(1)] = (\lambda + s)[F(t) - f(0)] + \mu t F'(t) - tF(t)$$

or

$$\mu(1-t)F'(t) = \{(\lambda + s) - \lambda t\}F(t) + \{\mu f(1) - (\lambda + s) f(0)\}$$

$$= \{(\lambda + s) - \lambda t\}F(t), \quad \text{because of (3.10.4)},$$

or

$$\frac{F'(t)}{F(t)} = \frac{(\lambda + s) - \lambda t}{\mu(1-t)} = \frac{\lambda}{\mu} + \frac{s}{\mu(1-t)}.$$

Integrating we get

$$\log F(t) = \log A + \frac{\lambda}{\mu}t - \frac{s}{\mu}\log(1-t) \quad \text{or}$$

$$F(t) = \frac{A\exp\left(\frac{\lambda t}{\mu}\right)}{(1-t)^{s/\mu}}.$$

Putting $t = 0$, $F(0) = A$, but $F(0) = f(0) = p_0^*(s)$, so that

$$F(t) = p_0^*(s)e^{\lambda t/\mu}(1-t)^{-s/\mu}. \tag{3.10.10}$$

Expanding $F(t)$ and comparing the coefficients of t^n, we get

$$p_n^*(s) = p_0^*(s)\sum_{j=0}^{n}\frac{\left(\frac{\lambda}{\mu}\right)^{n-j}}{(n-j)!}\frac{\left(\frac{s}{\mu}\right)\left(\frac{s}{\mu}+1\right)\ldots\left(\frac{s}{\mu}+j-1\right)}{j!}$$

$$= p_0^*(s)\sum_{j=0}^{n}\frac{\left(\frac{\lambda}{\mu}\right)^{n-j}\Gamma\left(\frac{s}{\mu}+j\right)}{(n-j)!\,j!\,\Gamma\left(\frac{s}{\mu}\right)}, \quad 0 \le n \le c-1, \ s \neq 0. \tag{3.10.11}$$

Now, we are to solve (3.10.6); putting $n = c$ and $n = c + r, r = 1, 2, \ldots$, we get

$$(\lambda + s + c\mu)\,p_c^*(s) - 1 = \lambda p_{c-1}^*(s) + c\mu p_{c+1}^*(s) \tag{3.10.12}$$

$$\text{and} \quad (\lambda + s + c\mu)\,p_{c+r}^*(s) = \lambda p_{c+r-1}^*(s) + c\mu p_{c+r+1}^*(s). \tag{3.10.13}$$

Denote $\sum_{r=0}^{\infty} p_{c+r}^* t^r = V(t)$. Multiplying (3.10.13) by t^r and adding for $r = 1, 2, 3, \ldots$, we get

$$(\lambda + s + c\mu)\sum_{r=1}^{\infty} p_{c+r}^*(s)t^r = \lambda t\sum_{r=1}^{\infty} p_{c+r-1}^*(s)t^{r-1} + \frac{c\mu}{t}\sum_{r=1}^{\infty} p_{c+r+1}^* t^{r+1}$$

or

$$(\lambda + s + c\mu)[V(t) - p_c^*(s)] = \lambda t V(t) + \frac{c\mu}{t}[V(t) - tp_{c+1}^*(s) - p_c^*(s)]$$

$$= \lambda t V(t) + \frac{c\mu}{t}[V(t) - p_c^*(s)]$$

$$+ \lambda p_{c-1}^*(s) - \{(\lambda + s + c\mu)\,p_c^*(s) - 1\},$$

(using (3.10.12)); or

$$[c\mu - (\lambda + s + c\mu)t + \lambda t^2]V(t) = p_c^*(s)(c\mu) + t\{-1 - \lambda p_{c-1}^*(s)\}.$$

Thus,

$$V(t) = \frac{c\mu p_c^*(s) - t\lambda p_{c-1}^*(s) - t}{c\mu - (\lambda + s + c\mu)t + \lambda t^2}. \tag{3.10.14}$$

To get $p_{c+r}^*(s)$ we have to expand the RHS of (3.10.14) in powers of t. Writing

$$[c\mu - (\lambda + s + c\mu)t + \lambda t^2] = \lambda(t - \alpha_1)(t - \alpha_2),$$

where

$$\begin{matrix}\alpha_1 \\ \alpha_2\end{matrix} = \frac{[(\lambda + s + c\mu) \pm \sqrt{\{(\lambda + s + c\mu)^2 - 4\lambda c\mu\}}]}{2\lambda},$$

$(\alpha_1, (\alpha_2)$ corresponding to the positive (negative) sign before the radical sign), we get

$$\begin{aligned}\frac{1}{c\mu - (\lambda + s + \mu)t + \lambda t^2} &= \frac{1}{\lambda(t - \alpha_1)(1 - \alpha_2)} \\ &= \frac{1}{\lambda(\alpha_1 - \alpha_2)}\left[\frac{1}{t - \alpha_1} - \frac{1}{t - \alpha_2}\right] \\ &= \frac{1}{\lambda(\alpha_1 - \alpha_2)}\left[\frac{-1}{\alpha_1}\left(1 - \frac{t}{\alpha_1}\right)^{-1} + \frac{1}{\alpha_2}\left(1 - \frac{t}{\alpha_2}\right)^{-1}\right] \\ &= \frac{1}{\lambda(\alpha_1 - \alpha_2)}\left[\frac{1}{\alpha_2}\sum_{k=0}^{\infty}\left(\frac{t}{\alpha_2}\right)^k - \frac{1}{\alpha_1}\sum_{k=0}^{\infty}\left(\frac{t}{\alpha_1}\right)^k\right].\end{aligned}$$

Thus,

$$\begin{aligned}p_{c+r}^*(s) &\equiv \text{coeff of } t^r \text{ in } V(t) \\ &\equiv \text{coeff of } t^r \text{ on the RHS of (3.10.14)} \\ &= \frac{1}{\lambda(\alpha_1 - \alpha_2)}\left[c\mu p_c^*(s)\left\{\frac{1}{\alpha_2^{r+1}} - \frac{1}{\alpha_1^{r+1}}\right\} - \lambda p_{c-1}^*(s)\left\{\frac{1}{\alpha_2^r} - \frac{1}{\alpha_1^r}\right\}\right. \\ &\quad \left. - \left(\frac{1}{\alpha_2^r} - \frac{1}{\alpha_1^r}\right)\right], \quad r = 0, 1, 2, \ldots \end{aligned} \tag{3.10.15}$$

Thus, we get all the coefficients $p_n^*(s), n > c$ in terms of $p_c^*(s)$ and $p_{c-1}^*(s)$. Again using (3.10.5), we get $p_c^*(s)$ in terms of $p_{c-1}^*(s)$ and $p_{c-2}^*(s)$ and all the $p_n^*(s), n \geq 1$ in terms of $p_0^*(s)$. The term $p_0^*(s)$ can be obtained by using the relation

$$\sum_{n=0}^{c-1} p_n^*(s) + \sum_{n=c}^{\infty} p_n^*(s) = \frac{1}{s},$$

or alternatively as follows. Write (3.10.5) as

$$(\lambda + s + c\mu)\, p_n^*(s) + (n-c)\mu p_n^*(s)$$
$$= \lambda p_{n-1}^*(s) + c\mu p_{n+1}^*(s) - (c-n-1)\mu p_{n+1}^*(s), \quad 1 \le n \le c-1. \tag{3.10.16}$$

Multiplying (3.10.16) by z^n for $1 \le n \le c-1$ and (3.10.6) by z^n for $n \ge c$ and adding for $n = 1, 2, \ldots$ to (3.10.4) and writing

$$P(z,s) = \sum_{n=0}^{\infty} p_n^*(s) z^s, \tag{3.10.17}$$

we get, on simplification,

$$P(z,s) = \frac{\mu(1-z)\sum_{n=0}^{c-1}(c-n)z^n p_n^*(s) - z^{c+1}}{\lambda z^2 - (\lambda + s + \mu)z + c\mu}. \tag{3.10.18}$$

Noting that $\lambda z^2 - (\lambda + s + \mu)z + c\mu = \lambda(z-\alpha_1)(z-\alpha_2)$, and considering that $P(z,s)$ exists in the unit circle, we find that the numerator of $F(z,s)$ must also vanish for $z = \alpha_2$. Thus, we have

$$\sum_{n=0}^{c-1}(c-n)\alpha_2^n p_n^*(s) = \frac{\alpha_2^{c+1}}{\mu(1-\alpha_2)}. \tag{3.10.19}$$

Now using the expression of $p_n^*(s), 0 \le n \le c-1$ as given in (3.10.11), we get

$$p_0^*(s) = \frac{\alpha_2^{c+1}}{\mu(1-\alpha_2)}\left[\sum_{n=0}^{c-1}(c-n)\alpha_2^n\left\{\sum_{j=0}^{n}\frac{\left(\frac{\lambda}{\mu}\right)^{n-j}\Gamma\left(\frac{s}{\mu}+j\right)}{(n-j)!\,j!\,\Gamma\left(\frac{s}{\mu}\right)}\right\}\right]^{-1}. \tag{3.10.20}$$

Thus, all $p_n^*(s)$ are known from(3.10.11), (3.10.15), and (3.10.20).

Note: Here we found the solution of the difference equations (3.10.4)–(3.10.6). For $n \ge c$, (3.10.6) is a difference equation of order 2 with constant coefficients and can be solved in the usual manner to get $p_n^*(s), n > c$. Equation (3.10.5) is a difference equation in $p_n^*(s)$ with coefficients as functions of n; a special technique is needed (as used here) to find $p_n^*(s), n \le c-1$. Then $p_0^*(s)$ can be obtained by either of the methods indicated. The special technique adopted here for the solution of (3.10.5) with (3.10.4) will be used subsequently in discussing a more complicated model $M/M/(1,b)/c$ (Section 4.6).

3.10.1.1 Steady-state distribution of the M/M/c model

Assume that $\rho = \lambda/c\mu < 1$. Then we can easily find

$$p_n = \lim_{t\to\infty} p_n(t) = \lim_{s\to 0} s p_n^*(s)$$

from (3.10.10), (3.10.15), and (3.10.20). From (3.10.8) and (3.10.10), we get

$$\begin{aligned}\lim_{s\to 0} s F(t) &= \lim_{s\to 0} \sum_{n=0}^{\infty} s p_n^*(s) t^n \\ &= \sum p_n t^n \\ &= p_0 e^{(\lambda/\mu)t} \\ &= \sum p_0 \left\{ \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \right\} t^n.\end{aligned}$$

Thus, for $0 \le n \le c-1$,

$$p_n = p_0 \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}. \tag{3.10.21}$$

From (3.10.5) for $n = c-1$

$$[\lambda + s + (c-1)\mu]\, p_{c-1}^*(s) = \lambda p_{c-2}^*(s) + c\mu p_c^*(s),$$

we get, multiplying both sides by s and taking limits as $s \to 0$,

$$[\lambda + (c-1)\mu]\, p_{c-1} = \lambda p_{c-2} + c\mu p_c.$$

Whence using (3.10.21), we get

$$\begin{aligned}p_c &= \frac{1}{c\mu} \left\{ [\lambda + (c-1)\mu] \left[\frac{\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!} \right] - \lambda \frac{\left(\frac{\lambda}{\mu}\right)^{c-2}}{(c-2)!} \right\} p_0 \\ &= \frac{1}{c\mu} \left[\frac{\lambda \left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!} \right] p_0 \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} p_0,\end{aligned}$$

so that $p_n = p_0(\lambda/\mu)^n/n!$ holds for $n = 0, 1, \ldots, c$. Now as $s \to 0$,

$$\alpha_1 \to \frac{[(\lambda + c\mu) + (\lambda - c\mu)]}{2\lambda} = 1$$

$$\text{and} \quad \alpha_2 \to \frac{[(\lambda + c\mu) + (\lambda - c\mu)]}{2\lambda} = \frac{1}{\rho}.$$

Thus, from (3.10.15), we get

$$\begin{aligned} p_{c+r} &= \lim_{s\to 0} s p^*_{c+r}(s) \\ &= \frac{1}{\lambda\left(1-\frac{1}{\rho}\right)}[c\mu p_c(\rho^{r+1}-1)] - \lambda p_{c-1}(\rho^r - 1)] \\ &= \rho^r p_c = \frac{\left(\frac{\lambda}{\mu}\right)^{c+r}}{c!c^r} p_0, \quad r = 0, 1, 2, \ldots, \quad \text{or} \\ p_n &= \frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}} p_0, \quad n = c, c+1, c+2, \ldots. \end{aligned} \tag{3.10.22}$$

From (3.10.21) and (3.10.22), using

$$\sum_{n=0}^{\infty} p_n = 1,$$

we get

$$\begin{aligned} p_0 &= \left[\sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \sum_{n=c}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{c!c^{n-c}}\right]^{-1} \\ &= \left[\sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(1-\rho)}\right]^{-1}. \end{aligned} \tag{3.10.23}$$

Thus, we get all $p_n, n = 0, 1, 2, 3, \ldots$ (as have been obtained in Section 3.6.1).

3.10.2 Busy period of an M/M/c queue

We consider here the approach by Chaudhry and Templeton (1973). A busy period of a multiserver queue may be defined as the interval of time commencing from the instant of arrival of a unit that makes a fixed number $k(\leq c)$ of channels busy to the first subsequent instant when the number of busy channels drops down to $(k-1)$. Let $\{N^*(t), t > 0\}$ be the stochastic process denoting the number of units present at the instant during the busy period T. This process avoids the states $0, 1, \ldots, k-1$—that is, the states are $k, k+1, \ldots$. Now $Pr\{N^*(0) = k\} = 1$, and the duration of the busy period is the interval $t(>0)$ for which $N^*(t)$ becomes $(k-1)$ for the first time. Let

$$q_n(t) = Pr\{N^*(t) = n \mid N^*(0) = k\}, \quad n \geq k,$$

and $b(t)$ be the PDF of the busy period T. Then it can be easily seen that

$$b(t) \equiv q'_{k-1}(t) = k\mu q_k(t).$$

Now q_n's satisfy the same equations as $p_n(t)$ in (3.10.2) restricted to $k \le n \le c-1$ and (3.10.3). That is,

$$q_n'(t) = -(\lambda + n\mu)q_n(t) + \lambda q_{n-1}(t) + (n+1)\mu q_{n+1}(t), \quad k \le n \le (c-1), \tag{3.10.24}$$

and $$q_n'(t) = -(\lambda + c\mu)q_n(t) + \lambda q_{n-1}(t) + c\mu q_{n+1}(t), \quad n \ge c, \tag{3.10.25}$$

where $\lambda q_{n-1}(t)$ is to be omitted from (3.10.24) when $n = k$ or from (3.10.25) when $n = k = c$, in which case (3.10.24) becomes redundant. Let $g_n(s)$ be the LT of $q_n(t)$. Then, taking the LT of (3.10.24) and (3.10.25), we get

$$(\lambda + s + n\mu)g_n(s) - 1 = \lambda g_{n-1}(s) + (n+1)\mu g_{n+1}(s), \quad k \le n \le c-1, \tag{3.10.26}$$

and $$(\lambda + s + c\mu)g_n(s) = \lambda g_{n-1}(s) + c\mu g_{n+1}(s), \quad n \ge c. \tag{3.10.27}$$

It is to be noted that (-1) in the LHS of (3.10.26) is to be used only when $n = k$. Further, when $n = k = c$, then (3.10.26) becomes redundant and (3.10.27) will have a term (-1) on the LHS of (3.10.27). Define

$$V(z, s) = \sum_{n=k}^{\infty} g_n(s) z^n.$$

Multiplying (3.10.26) and (3.10.27) by z^n and adding for appropriate values of n, we get

$$(\lambda + s + c\mu)V(z, s) - \sum_{n=k}^{c-1} \mu(c-n)g_n(s)z^n - z^k$$

$$= \lambda z V(z, s) + \frac{c\mu}{z} V(z, s) + \sum_{n=k}^{c-1} (n + 1 - c)\mu g_{n+1}(s) z^n,$$

which can be written as

$$V(z, s) = \frac{z^{k+1} + \mu(z-1)\sum_{n=k}^{c-1}(c-n)g_n(s)z^n - k\mu g_k(s)z^k}{-\lambda z^2 + (\lambda + s + c\mu)z - c\mu}. \tag{3.10.28}$$

The denominator put to zero

$$\lambda z^2 - (\lambda + s + c\mu) + c\mu = 0$$

has two roots, α_1 and α_2,

$$\frac{\alpha_1}{\alpha_2} = \frac{\lambda + s + c\mu \pm \sqrt{(\lambda + s + c\mu)^2 - 4\lambda c\mu}}{2\lambda}$$

of which α_2 is of modulus less than 1. Considering that $V(z,s)$ exists in the unit circle, we see that the numerator of (3.10.28) must also vanish for $z = \alpha_2$. Putting $z = \alpha_2$ in the numerator of (3.10.28) equated to zero, we get

$$\alpha_2^{k+1} + \mu(\alpha_2 - 1)\sum_{n=k}^{c-1}(c-n)g_n(s)\alpha_2^n - k\mu g_k(s)\alpha_2^k = 0 \quad \text{or}$$

$$(1-\alpha_2)\sum_{n=k}^{c-1}(c-n)g_n(s)\alpha_2^n + kg_k(s)\alpha_2^k = \frac{(\alpha_2^{k+1})}{\mu}. \tag{3.10.29}$$

The equation involves $(c-k)$ unknowns $g_n(s)$, $(k \le n \le c-1)$. These $(c-k)$ unknowns can be determined from (3.10.29) and (3.10.26). Thus, $g_k(s)$ can be found; its inversion gives $g_k(t)$. Note that (3.10.29) is sufficient to determine $g_k(s)$ in the cases $k = c$ or $k = (c-1)$. When $k = c$, the term $\sum_k^{c-1}$ will not occur.

3.10.2.1 Particular cases

I. $k = c$—that is, the busy period starts from the instant when all the servers get busy to the instant when one of the servers becomes free for the first time.

Putting $k = c$ in (3.10.29) we get

$$c\mu g_c(s) = \alpha_2.$$

Inverting the LT we get

$$b(t) = c\mu q_c(t) = \left[\frac{e^{-(\lambda+c\mu)t}}{t}\right]\sqrt{\frac{c\mu}{\lambda}}I_1(2t\sqrt{c\lambda\mu})$$

$$= \left(\frac{1}{t}\right)\rho^{-1/2}e^{-(\lambda+c\mu)t}I_1(2t\sqrt{c\lambda\mu}). \tag{3.10.30}$$

We have

$$E(T) = \frac{d}{ds}b^*(s)|_{s=0} = \frac{1}{c\mu-\lambda}.$$

Comparing the corresponding result for the single channel case (corresponding to $c = 1$) given in (3.9.25), we see the busy period distribution for the c-channel case (when the busy period is defined as the period during which all the servers remain busy—that is, $k = c$) can be obtained form the single-channel case by replacing μ by $c\mu$. This should be intuitively clear as the whole set of c-servers can then be considered as a compact set of "a single server" with rate $c\mu$.

II. $k = c - 1$.

Then from (3.10.29), we get

$$(1-\alpha_2)g_{c-1}(s)\alpha_2^{c-1} + (c-1)g_{c-1}(s)\alpha_2^{c-1} = \frac{\alpha_2^c}{\mu} \tag{3.10.31}$$

$$\text{or} \quad \mu g_{c-1}(s) = \frac{\alpha_2}{c-\alpha_2}$$

$$= \frac{c\mu}{\lambda}\frac{1}{\alpha_1}\left(1 - \frac{c\mu}{\lambda}\frac{1}{\alpha_1}\right)^{-1}$$

$$= \sum_{n=0}^{\infty}\left(\frac{\mu}{\lambda}\right)^{n+1}\frac{1}{\alpha_1^{n+1}},$$

$$\text{since} \quad \left|\frac{\mu}{\lambda}\frac{1}{\alpha_1}\right| = \left|\left(\frac{1}{c}\right)\alpha_2\right| < 1.$$

$$b(t) = (c-1)\mu g_{c-1}(t)$$

$$= \frac{c-1}{t}e^{-(\lambda+c\mu)t}\sum_{n=0}^{\infty}(n+1)\rho^{(-1/2)(n+1)}I_{n+1}(2t\sqrt{c\lambda\mu}). \tag{3.10.32}$$

Inverting the LT, we get

In particular, when $c = 2, k = 1$, we get from (3.10.32)

$$b(t) = \frac{1}{t}e^{-(\lambda+2\mu)t}\sum_{n=0}^{\infty}(n+1)\rho^{(-1/2)(n+1)}I_{n+1}(2t\sqrt{2\lambda\mu}) \tag{3.10.33}$$

(where $\rho = \lambda/2\mu$).

3.10.3 Transient-state distribution of the output of an $M/M/c$ queue

In Section 3.6.4 we saw that under steady state the output process of an $M/M/c$ queueing system is Poisson, with the same rate as that of the input Poisson process. In Section 3.9.1 we discussed the transient-state properties of the $M/M/1$ queueing system. We also discussed the importance of the transient-state results. The transient-state properties of the output of an $M/M/c$ queue with c-servers have been investigated by Everitt and Downs (1984). We discuss their approach next.

Let λ and μ be the arrival and service rates, respectively, of an $M/M/c$ system. Let $N(t)$ denote the number of customers in the system at instant t, $A(t)$ denote the number of arrivals in $(0, t)$, and $D(t)$ denote the number of departures in $(0, t)$, so that

$$N(t) = N(0) + A(t) - D(t). \tag{3.10.34}$$

Now it can be easily seen that $\{N(t), t > 0\}$ is discrete state space and continuous-time Markov process. Let

$$\begin{aligned} Pr\{D(t) = k, A(t) = n \mid N(0) = m\} \\ = P_t\{k, n \mid m\}, \end{aligned} \tag{3.10.35}$$

$$\begin{aligned} Pr\{D(t) = k \mid A(t) = n, N(0) = m\} \\ = P_t\{k \mid n, m\}, \end{aligned} \tag{3.10.36}$$

and $Pr\{A(t) = n \mid N(0) = m\} = P_t(n \mid m)$;

then

$$\begin{aligned} P_t\{k \mid n, m\} &= \frac{P_t\{k, n \mid m\}}{P_t\{n \mid m\}} \\ &= \frac{P_t\{k, n \mid m\}}{\frac{e^{-\lambda t}(\lambda t)^n}{n!}} \\ &= \frac{e^{\lambda t}\lambda^{-n} P_t\{k, n \mid m\}}{\left(\frac{t^n}{n!}\right)}. \end{aligned} \tag{3.10.37}$$

Denote

$$\begin{aligned} R_t(k \mid n, m) &= \frac{t^n}{n!} P_t(k \mid n, m) \\ &= \lambda^{-n} e^{\lambda t} P_t(k, n \mid m). \end{aligned} \tag{3.10.38}$$

Theorem 3.3. *The probabilities $P_t(k, n \mid m)$ and $R_t(k \mid n, m)$ satisfy the following differential-difference equations:*

$$\begin{aligned} \frac{d}{dt} P_t(k, n \mid m) &= -\{\lambda + \mu \min(c, m + n - k)\} P_t(k, n \mid m) + P_t(k, n - 1 \mid m) \\ &\quad + \mu \min(c, m + n - k - 1) P_t(k - 1, n \mid m) \quad \textit{and} \end{aligned} \tag{3.10.39}$$

$$\begin{aligned} \frac{d}{dt} R_t(k \mid n, m) &= -\mu \min(c, m + n - k) R_t(k \mid n, m) + R_t(k \mid n - 1, m) \\ &\quad + \mu \min(c, m + n - k - 1) R_t(k - 1 \mid n, m). \end{aligned} \tag{3.10.40}$$

Proof: When the number of arrivals and departures are r and s, respectively, at any time, the number of servers at that time is $\min(c, m + r - s)$, and so the rate of service is $\mu \min(c, m + r - s)$. Using this fact, it can be seen that forward Chapman-Kolmogorov equations lead to the relation (3.10.39).

$$\text{Writing} \quad P_t(k, n \mid m) = \lambda^n e^{-\lambda t} R_t(k \mid n, m)$$

and putting the expressions in terms of R_t in (3.10.40) and on simplifying, we at once get (3.10.40). ■

Note: The two relations (3.10.39) and (3.10.40) are equivalent. We get the behavior of the output process from both of them. But the relation (3.10.40) is simpler as it holds for all arrival rates λ (3.10.40 is independent of λ).

The general solution of (3.10.40) in time domain is difficult. A solution in terms of LST of R_t is given by Everitt and Downs (1984). (See Problems and Complements 3.16.)

3.11 Multichannel Queue with Ordered Entry

Two assumptions in usual multichannel queueing problems are as follows.

(1) There is a single queue or waiting line for all the channels.

(2) Any unit who finds on entering the system that more than one service channel is free (or server idle) chooses a channel at random.

There is a class of problems, with practical applications, where neither of these two assumptions holds. Here each channel has its own queue or waiting line, and the entering unit cannot choose the channel at random but rather must use the first channel it comes to. We number the service channels $1, 2, \ldots, n$. The entering unit tries channel 1 first, and if it is free, he must go into it; if channel 1 is busy, the arriving unit joins the queue in front of channel 1 unless the queue already there reaches the maximum queue length that can be accommodated there. In other words, if the maximum queue length is $(M - 1)$ in channel 1 so that the total capacity in channel 1 is M with the unit in service, and if the entering unit finds fewer than M in the system in channel 1, he joins channel 1, whereas if he finds M in the system in channel 1, he tries the next channel in order—that is, channel 2. In this way each arriving unit proceeds to test each service facility in order until he finds *one* channel with a number in the system for that channel less than the capacity of that channel. An arriving unit examines each channel from 1 to n in order; if he finds that each channel has reached the maximum capacity for that channel, then the unit leaves and is lost to the system.

This is a *multiqueue* problem with one separate queue before each channel and with *ordered entry.*

The queueing behavior in a supermarket is more or less similar to this problem. This model is useful in studying conveyor systems in industrial engineering. Disney (1962, 1963) first considered this model, We follow his approach here.

We make the following assumptions.

(1) Units arrive into the system from an infinite Poisson source with rate λ—that is, arrivals are in accordance with a Poisson process with parameter λ.

(2) Service-time distribution in each channel is exponential with parameter μ.

(3) Each service facility contains one server, and the queue discipline in each channel is FCFS.

We confine ourselves here to steady-state analysis.

3.11.1 Two-channel model with ordered entry (with finite capacity)

Let M and N be the maximum capacity (for the queue and the unit under service) of channels 1 and 2, respectively. Assume that the system is in steady state. Let (i, j) be the state of the system where i and j denote the number in channel 1 and channel 2, respectively. Each value i, j includes the unit in service, if any, $0 \leq i \leq M, 0 \leq j \leq N$, $p_{ij} = Pr\{\text{system is in state } (i, j)\}$.

The difference equations can be written by using the rate equality principle. Consider the state $(0,0)$. It can leave state $(0,0)$ through an arrival to channel 1 when the state becomes $(1,0)$. Thus, the rate of leaving state $(0,0)$ is $\lambda p_{0,0}$. It can enter the state $(0,0)$ either from the state $(1,0)$ through service completion of the unit under service in channel 1 or from the state $(0,1)$ through service completion of the unit under service in channel 2, the rate of entering state $(0,0)$ being $\mu p_{1.0} + \mu p_{0.1}$. Thus,

$$\lambda p_{0.0} = \mu p_{1.0} + \mu p_{0.1}. \tag{3.11.1}$$

Consider the state $(M, 0)$. It can leave this state though a service completion or through an arrival of a unit that will join channel 2 as channel 1 is full. Thus, the rate of leaving the state $(M, 0)$ is $\mu p_{M,0} + \lambda p_{M,0}$. It can enter the state $(M, 0)$ from either the state $(M - 1, 0)$ through an arrival or from the state $(M, 1)$ through a departure, the rate being $\lambda p_{M-1,0} + \mu p_{M,1}$. Thus,

$$\lambda p_{M,0} + \mu p_{M,0} = \lambda p_{M-1,0} + \mu p_{M,1}. \tag{3.11.2}$$

Consider the (M, N). It can leave that state through a departure either from channel 1 or from channel 2, the rate of leaving being $2\mu p_{M,N}$. It can enter the state (M, N) either from the state $(M - 1, N)$ or from the state $(M, N - 1)$ through an arrival, the rate of entering being $\lambda p_{M-1,N} + \lambda p_{M,N-1}$. Thus,

$$2\mu p_{M,N} = \lambda p_{M-1,N} + \lambda p_{M,N-1}. \tag{3.11.3}$$

Consider the state $(0, N)$. It can leave that state either through an arrival to the channel 1 or from a departure from the channel 2, the rate being $(\lambda + \mu) p_{0,N}$. It can enter that state only from the state $(1, N)$ through a departure from the channel 1, the rate being $\mu p_{1,N}$. Thus,

$$(\lambda + \mu) p_{0,N} = \mu p_{1,N}. \tag{3.11.4}$$

Consider the state $(M, L), 0 < L < N$. It can leave the state (M, L) either from the state $(M, L - 1)$ through an arrival who joins channel 2 or through a departure from either of the two channels, the rate being $(\lambda + 2\mu) p_{M,L}$. It can enter the state (M, L) from the state $(M - 1, L)$ through an arrival or from the state $(M, L - 1)$ through an arrival, or from the state $(M, L + 1)$ through

a departure, the rate being $\lambda p_{M-1,L} + \lambda p_{M,L-1} + \mu p_{M,L+1}$. Thus,

$$(\lambda + 2\mu)\, p_{M,L} = \lambda p_{M-1,L} + \lambda p_{M,L-1} + \mu p_{M,L+1}. \tag{3.11.5}$$

Consider the state $(K, N), 0 < K < M$. It can leave that state through an arrival who joins channel 1 or through a departure from either of the two channels, the rate being $(\lambda + 2\mu)\, p_{K,N}$. It can enter the state (K, N) either from the state $(K-1, N)$ through an arrival or from the state $(K+1, N)$ through a departure, the rate being $\lambda p_{K-1,N} + \mu p_{K+1,N}$. Thus,

$$(\lambda + 2\mu)\, p_{K,N} = \lambda p_{K-1,N} + \mu p_{K+1,N}. \tag{3.11.6}$$

Consider the state $(0, L), 0 < L < N$. It can leave that state either through an arrival (to channel 1) or from a departure (from channel 2), the rate being $(\lambda + \mu)\, p_{0,L}$. It can enter that state $(0, L)$ from the state $(1, L)$ or from the state $(0, L + 1)$ through a departure, the rate being $\mu p_{1,L} + \mu p_{0,L+1}$. Thus,

$$(\lambda + \mu)\, p_{0,L} = \mu(p_{1,L} + p_{0,L+1}). \tag{3.11.7}$$

Consider the state $(K, 0), 0 < K < M$. It can leave that state through an arrival (to channel 1) or a departure (from channel 1), the rate being $(\lambda + \mu)\, p_{K,0}$. It can enter the state $(K, 0)$ either from the state $(K - 1, 0)$ through an arrival or from the states $(K + 1, 0)$ or $(K, 1)$ through a departure, the rate being $\lambda p_{K-1,0} + \mu p_{K+1,0} + \mu p_{K,1}$. Thus,

$$(\lambda + \mu)\, p_{K,0} = \mu p_{K-1,0} + \mu p_{K+1,0} + \mu p_{K,1}. \tag{3.11.8}$$

Finally, consider the state $(K, L), 0 < K < M, 0 < L < N$. It can leave that state either through an arrival (to channel 1) or from a departure from either of the two channels, the rate being $(\lambda + 2\mu)\, p_{K,L}$. It can enter that state either from the state $(K - 1, L)$ through an arrival or from either of the states $(K + 1, L)$ or $(K, L + 1)$ through a departure, the rate being $\lambda p_{K-1,L} + \mu(p_{K+1,L} + p_{K,L+1})$. Thus,

$$(\lambda + 2\mu)\, p_{K,L} = \lambda p_{K-1,L} + \mu(p_{K+1,L} + p_{K,L+1}). \tag{3.11.9}$$

These nine sets of equations give the complete description of the system. Out of the preceding set of nine equations, one is dependent, implying that we solve all others and get p_{ij} in terms of one particular p_{kl}. Then using the normalizing condition $\sum_{i,j} p_{ij} = 1$, we can find this particular p_{kl} so that all p_{ij} are completely determined. However, the equations of the system cannot be solved recursively as is usually done.

3.11.2 The case $M = 1, N = N$

Consider the two-channel model with ordered entry and with no waiting space before channel 1 and with $N - 1$ waiting spaces before channel 2. Then we put $M = 1$ in the relevant equations. From (3.11.1), (3.11.4), and (3.11.7) and

writing $\rho = \lambda/\mu$, we get

$$\begin{aligned} -\rho p_{0,0} + p_{0,1} + p_{1,0} &= 0 \\ -(1+\rho)\, p_{0,L} + p_{0,L+1} + p_{1,L} &= 0, \quad 0 < L < N, \\ -(1+\rho)\, p_{0,N} + p_{1,N} &= 0. \end{aligned} \tag{3.11.10}$$

Denote

$$\boldsymbol{P}_0 = \begin{pmatrix} p_{00} \\ \vdots \\ p_{0i} \\ \vdots \\ p_{0N} \end{pmatrix}, \qquad \boldsymbol{P}_1 = \begin{pmatrix} p_{10} \\ \vdots \\ p_{1j} \\ \vdots \\ p_{1N} \end{pmatrix}. \tag{3.11.11}$$

Writing

$$\boldsymbol{A}_{11} = \begin{pmatrix} -\rho & 1 & 0 & \cdots & 0 \\ 0 & -(1+\rho) & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & -(1+\rho) \end{pmatrix}, \tag{3.11.12}$$

we can express the preceding equations as

$$\boldsymbol{A}_{11}\boldsymbol{P}_0 + \boldsymbol{P}_1 = \boldsymbol{O}. \tag{3.11.13}$$

Putting $M = 1, N = N$ in Equations (3.11.2), (3.11.3), and (3.11.5), we get

$$\begin{aligned} -(1+\rho)\, p_{1,0} + p_{1,1} + \rho p_{0,0} &= 0 \\ p_{1,L-1} - (\rho+2)\, p_{1,L} + p_{1,L+1} + p_{0,L} &= 0, \quad 0 < L < N, \\ \rho p_{1,N-1} - 2\, p_{1,N} + \rho p_{0,N} &= 0. \end{aligned} \tag{3.11.14}$$

Writing

$$\boldsymbol{A}_{33} = \begin{pmatrix} -(1+\rho) & 1 & 0 & \cdots & 0 \\ \rho & -(2+\rho) & 1 & \cdots & 0 \\ 0 & 0 & -(2+\rho) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & -2 \end{pmatrix}, \tag{3.11.15}$$

we can express the preceding Eqs. (3.11.14) as

$$\boldsymbol{A}_{33}\boldsymbol{P}_1 + \rho \boldsymbol{I}\boldsymbol{P}_0 = \boldsymbol{O}, \tag{3.11.16}$$

where $\boldsymbol{I}$ is the unit matrix of order $(N \times N)$. Equations (3.11.13) and (3.11.16) describe the system. From Eq. (3.11.13) we can immediately see that

$$\boldsymbol{P}_1 = -\boldsymbol{A}_{11}\boldsymbol{P}_0. \tag{3.11.17}$$

This relation allows $\boldsymbol{P}_1$ to be expressed in terms of $\boldsymbol{P}_0$.

Substituting in Eq. (3.11.16), we get

$$\begin{aligned} A_{33}(-\boldsymbol{A}_{11}\boldsymbol{P}_0) + \rho\boldsymbol{I}\boldsymbol{P}_0 = \boldsymbol{O} \quad &\text{or} \\ (-\boldsymbol{A}_{33}\boldsymbol{A}_{11} + \rho\boldsymbol{I})\boldsymbol{P}_0 = \boldsymbol{O}, \quad &(3.11.18) \end{aligned}$$

which is a set of homogeneous equations in terms of the unknown $\boldsymbol{P}_0$. The coefficient matrix is of rank N, and hence this set of homogeneous equations can be solved, and all terms of $\boldsymbol{P}_0$ can be solved in terms of any one of them. Using the relation

$$\sum_{i=0}^{1}\sum_{j=0}^{N} p_{ij} = 1, \tag{3.11.19}$$

we can find this term. Hence, the set of equations can be completely solved.

3.11.3 Particular case: $M = N = 1$ (overflow system)

Consider the two-channel model with ordered entry having no space for queue before any of the servers (except the ones in service, if any). Then putting $M = N = 1$ in Eqs. (3.11.1)–(3.11.4), we shall get the following equations

$$\lambda p_{0,0} = \mu p_{1,0} + \mu p_{0,1} \tag{3.11.20}$$

$$(\lambda + \mu)\, p_{1,0} = \lambda p_{0,0} + \mu p_{1,1} \tag{3.11.21}$$

$$2\mu p_{1,1} = \lambda p_{0,1} + \lambda p_{1,0} \tag{3.11.22}$$

$$(\lambda + \mu)\, p_{0,1} = \mu p_{1,1}. \tag{3.11.23}$$

We can use the matrix method outlined previously. Here we solve them recursively. Denote $\rho = \lambda/\mu$. From (3.11.23) we get

$$p_{1,1} = (1 + \rho)\, p_{0,1}. \tag{3.11.24}$$

Putting this value of $p_{1,1}$ in (3.11.22), we get

$$p_{1,0} = \left(\frac{1}{\rho}\right)[2(1 + \rho) - \rho]\, p_{0,1} = \left(\frac{2 + \rho}{\rho}\right) p_{0,1}. \tag{3.11.25}$$

From (3.11.21) we then get

$$p_{0,0} = \left(\frac{1}{\rho}\right)[(1+\rho)p_{1,0} - p_{1,1}]$$
$$= \frac{2(1+\rho)}{\rho^2} p_{0,1}. \tag{3.11.26}$$

Equation (3.11.20) is the dependent equation. Using $\sum_{i,j=0}^{1} p_{i,j} = 1$, we get from Eqs. (3.11.24)–(3.11.26),

$$p_{0,1}\left[1 + (1+\rho) + \frac{2+\rho}{\rho} + \frac{2(1+\rho)}{\rho^2}\right] = 1 \quad \text{or} \tag{3.11.27}$$
$$p_{0,1} = \frac{\rho^2}{(\rho+1)(\rho^2+2\rho+2)}$$

and so

$$p_{1,0} = \frac{\rho(2+\rho)}{(\rho+1)(\rho^2+2\rho+2)},$$
$$p_{1,1} = \frac{\rho^2}{(\rho^2+2\rho+2)}, \quad \text{and} \tag{3.11.28}$$
$$p_{0,0} = \frac{2}{(\rho^2+2\rho+2)}.$$

If N denotes the total number in the system (in both the channels combined), then

$$Pr(N=0) = p_{0,0} = \frac{1}{\sum_{r=0}^{2} \frac{\rho^r}{r!}}.$$
$$Pr(N=1) = p_{1,0} + p_{0,1} = \frac{\rho}{\sum_{r=0}^{2} \frac{\rho^r}{r!}}, \quad \text{and} \tag{3.11.29}$$
$$Pr(N=2) = p_{1,1} = \frac{\rho^2/2}{\sum_{r=0}^{2} \frac{\rho^r}{r!}},$$

as is expected, these being the corresponding probabilities for the system $M/M/2/2$. Here $p_{1,1}$ is the probability that both the channels are busy—that is, the probability that an arriving unit will be lost to the system. This gives the Erlang's loss formula for $M/M/2/2$.

The expected total number in the system equals

$$\sum_{k=0}^{2} kPr(N=k) = \frac{\rho(\rho+1)}{\sum_{r=0}^{2} \frac{\rho^r}{r!}}. \tag{3.11.30}$$

Disney (1962, 1963) discusses further analysis of the working of such a system.

Notes:

(1) The system is called an overflow network system. (See the chapter on networks.) The sojourn times in such a system are of simple structure and can be found trivially. For example, if S_m is the stationary sojourn time of the mth customer, then

$$Pr(S_m = 0) = p_{1,1} = \frac{\rho^2}{(\rho^2 + 2\rho + 2)}$$

$$Pr(S_m \leq t) = \frac{2(1+\rho)(1 - e^{\mu t})}{(\rho^2 + 2\rho + 2)}.$$

(2) Consider an *overflow system*: a two-channel ordered-entry model with no waiting space, $M = N = 1$.

If $t_{m,j}$ is the time of the mth overflow from server j $(j = 1, 2)$, then in equilibrium, the *interoverflow time* $\{t_{m+1,j} - t_{m,j}, m = 1, 2, \ldots\}$ is a sequence of IID random variables. The equilibrium overflow (interoverflow) distribution function $\Phi_j(x)$, from the jth server, $j = 1, 2$, satisfies the intergral equation (Palm's equation)

$$\Phi_j(x) = \int_0^x e^{-\mu_j v} d\Phi_{j-1}(v) + \int_0^x (1 - e^{-\mu_j v})\Phi_j(x - v) d\Phi_{j-1}(v)$$

with

$$\Phi_0(x) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

The overflow distribution for the first server is given by

$$\Phi_1(x) = ae^{-bx} + (1 - a)e^{-cx},$$

where a, b, and c are functions of λ, μ_i (the rate of the ith server, $i = 1, 2$). The preceding shows that the overflow distribution is a sum of exponential terms and that the overflow process is not Poisson but is a renewal process.

The overflow networks first appeared in telephony. Palm (1943) considered overflow distribution. The integral equation involving $\Phi_j(x)$ is known as *Palm's integral equation.* Khinchin (1960), considering Palm's treatment, introduced the idea of Palm functions into modern point process theory.

3.11.4 Output process

If $t^0_{m,1}$ is the time of the mth output from server 1, then in equilibrium $\{t^0_{m+1,1} - t^0_{m,1}, m = 1, 2, \ldots\}$ is a sequence of IID random variables whose distribution is the sum of two independent RVs: the service-time distribution and the interarrival-time distribution. (See Disney and König (1985).)

Remarks: The queueing system with ordered entry has received considerable attention because of its importance in application, mainly in conveyor theory. (See Muth and White (1979) for a survey.) For further work in this area, reference may be made, for example, to Elsayed (1983), Elsayed and Elayat (1976), Elsayed and Proctor (1977), Gregory and Litton (1975), Lin and Elsayed (1978), Matsui and Fukuta (1977), Nawijn (1983, 1984), Newell (1984), Pourbabai (1987), Pourbabai and Sonderman (1986), Pritsker (1966), Proctor *et al.* (1977), Sonderman (1982), Yao (1987), and Shanthikumar and Yao (1987).

Apart from conveyor systems, the ordered entry model also applies to communication networks, such as System Network Architecture (SNA) (see, for example, Gray and Mcneill (1979)). A third area of application is database systems—see, for example, Cooper and Solomon (1984).

Problems and Complements

3.1. (a) Combination of service channels. Consider that two identical $M/M/1$ queueing systems with the same rates λ, μ (intensity $\rho = \lambda/\mu$) are in operation side by side (with separate queues) in a premises. Show that the distribution of the total number N in the two systems taken together is given by

$$Pr(N = n) = (n + 1)(1 - \rho)^2\rho^n, \quad n \geq 0.$$

(b) Consider (i) c number of $M/M/1$ queues each with rates λ and μ, (ii) a standard $M/M/c$ queue with rates $c\lambda$ and μ. Show that the average waiting time is less for the system (ii) for any $c > 1$. (This provides justification for having a common queue for a multiserver system in a bank, reservation counter, and so on.) (See Smith and Whitt (1981) and Rothkopf and Rech (1987).)

3.2. Dufkova and Zitek's (1975) model of an $M/M/1$ system with a class of queueing disciplines.
Suppose that when the server becomes free, he accepts either the first in the queue with probability δ or the last in the queue with probability $1 - \delta$, so that $\delta = 1$ implies FCFS and $\delta = 0$ LCFS disciplines. Show the LST of the waiting-time distribution is

$$w_q^*(s) = (1 - \rho) + \rho\left[1 - \frac{sR}{\lambda(1 - R)}\right],$$

where $R = R(\delta)$ is the unique real root in (0, 1) of the equation

$$(1 - \delta)\mu x^2 - (\mu + \lambda - \delta\lambda + s)x + \lambda = 0.$$

Show that the mean waiting time

$$E(W_q) = \frac{\lambda}{\mu(\mu - \lambda)}$$

is independent of R and of δ (as can be expected, it is independent of the queue discipline), and

$$\text{var}(W_q) = \frac{2\lambda}{(\mu - \lambda)^2(\mu - \lambda + \delta\lambda)} - \frac{\rho^2}{(\mu - \lambda)^2},$$

and the exterme values of var (W_q) are for $\delta = 1$ and $\delta = 0$. Show that the results also hold good for the $M/M/c$ model.

3.3. An $M/M/1$ queue with control-limit policy and exponential start-up time (Baker, 1973).

Here the control policy is to turn off the system and withdraw the server when the system becomes empty and to turn on the system when the system size reaches $n(>0)$. When the system is turned on, it cannot immediately serve customers but requires some time to start up. It is assumed that the time required for start-up is exponential with mean $1/\gamma$. The interarrival and service-time distributions are exponential with means $1/\lambda$ and $1/\mu$, respectively. Suppose that $p_{i,s}$ denotes the steady-state probability that there are i customers in the system and the server state is s (where $s = 0$ implies an idle state of the server in which the start-up has not begun or has not been completed) and $s = 1$ denotes the busy state of the server (in which service is being performed). Suppose that $P_i = P_{i,0} + P_{i,1}$ is the steady-state probability that there are i in the system.

Denote

$$\rho = \frac{\lambda}{\mu}(<1),$$

$$\theta = \frac{\lambda}{(\lambda + \gamma)}, \quad \text{and}$$

$$\omega = \frac{\gamma}{\mu}.$$

Show that the system satisfies the set of equations

$$P_{1,1} = \rho P_{0,0},$$

$$P_{i,0} = P_{i-1,0}, \quad 1 \le i \le n - 1,$$

$$P_{i,0} = \theta P_{i-1,0}, \quad n \le i,$$

$$P_{i,1} = (1 + \rho) P_{i-1,1} - \rho P_{i-2,1}, \quad 2 \le i \le n,$$

$$P_{i,1} = (1 + \rho) P_{i-1,1} - \rho P_{i-2,1} - \omega P_{i-1,0}, \quad n + 1 \le i.$$

Show further that

$$P_i = \alpha(1-\rho^{i+1}), \quad 0 \le i \le n-1,$$

$$P_i = \frac{\alpha[(1-\rho^{n-1})\rho^{i-n+2} + (1-\rho)(\rho^{i-n+2} - \theta^{i-n+2})]}{(\rho-\theta)}, \quad n \le i,$$

where $\rho \neq \theta$, and $\alpha = (1-\theta)/[\theta + n(1-\theta)]$. Further show that the mean number in the system $E(N)$ equals

$$E(N) = \sum_{i=0}^{\infty} i P_i = \frac{\alpha n(n-1)}{2} + \frac{(\rho - 2\rho\theta + \theta)}{(1-\rho)(1-\theta)}.$$

See Borthakur *et al.* (1987) for a more general model, and Böhm and Mohanty (1990) for transient solution (through discrete-time analogue and use of combinatorial arguments) of $M/M/1$ queue under control limit policy and zero start-up time. See also Lee (1990).

3.4. Naor's model for regulation of queue size (Naor, 1969).
Suppose that the cost to a customer of staying in a queue (i.e., for queueing) is c per unit time and that the reward collected at the end of service is R. Because of these costs a newly arrived customer weighs two alternatives: to join or not to join the queue by the net gains associated with them. For "individual optimization," show that the critical number n_s of customers is given by

$$n_s = \left[\frac{R\mu}{c}\right],$$

where $[\alpha]$ is the largest integer not exceeding α. For overall or "collective optimization," show that the critical number n_0 satisfies

$$\frac{n_0(1-\rho) - \rho(1-\rho^{n_0})}{(1-\rho)^2} \le \frac{R\mu}{c} < \frac{(n_0+1)(1-\rho) - \rho(1-\rho^{n_0-1})}{(1-\rho)^2}.$$

That is, n_0 equals

$$n_0 = [v_0],$$

where v_0 is given by

$$\frac{v_0(1-\rho) - \rho(1-\rho^{v_0})}{(1-\rho)^2} = \frac{R\mu}{c}.$$

Further, show that $v_0 \le R\mu/c$, where the sign of equality holds if $R\mu/c$ equals unity. (λ, μ, and ρ denote, respectively, the mean arrival rate, mean service rate, and traffic intensity in an $M/M/1$ queue with balking.)

3.5. De Vany's Model (1976).
De Vany uses the $M/M/1/K$ model to determine the effective demand function under the conditions: (1) the arrival stream is Poisson with rate $\lambda(p)$ (where p is price), (2) customers' orders are serviced on an FCFS basis with exponentially distributed service time with parameter μ, and (3) customers have a common balking value K that depends on the expected benefits of purchase at the firm relative to the expected benefits of purchase from an alternative supplier. Show that the effective demand for service equals

$$\begin{aligned}\lambda' &= \lambda[1 - p_k] \\ &= \lambda(p)[1 - B(\lambda(p), \mu, K)],\end{aligned}$$

where $B = p_k$ is the balking probability. Note that $\lambda' < \lambda$; this implies that there is a kind of excess demand of the firm's product in the sense that some potential customers arrive but leave to go to the alternative firm as the queue they find on arrival is at the balking length K.
The equilibrium mean rate of demand λ' is a function of price, mean capacity μ, and the balking value K.

Taking derivatives show that the price affects both the arrival rate and the proportion of those who stay on. A higher price reduces both the arrival rate and the balking value K, which causes a greater proportion of arrivals to balk.
Show that the demand curve is less elastic than the potential curve.

3.6. Show that for the $M/M/1$ queueing system starting with k customers at time 0, the joint distribution of the busy period T_k, and the number $N(T_k)$ served during the busy period initiated by k customers is given by

$$P\{t \leq T_k < t + dt, N(T_k) = n\} = e^{-(\lambda+\mu)t} \frac{k\lambda^{n-k}\mu^n t^{2n-k-1}}{n!(n-k)!} dt.$$

Hence, find the marginal distributions, that is, the PDF for T_1 (busy period initiated by a single customer) and the distribution of $N(T_1) = N$ (Prabhu, 1965).

3.7. Show that the expected busy period for an $M/M/1/K$ queueing system equals

$$E(T) = \frac{1 - a^{K+1}}{\mu(1-a)}, \quad \text{for } a = \frac{\lambda}{\mu} \neq 1.$$

Show further that the expected number of loss during a busy period is $<, =$, or > 1 according as a is $<, =$ or > 1.
The result holds for general service time as well (Problems and Complements 6.11).

(Hints: Use $E(T)/E(I) = (1 - p_0)/p_0$. Here $E(I) = 1/\lambda'$ where λ' is the mean *effective* arrival rate.)

3.8. Show that for the $M/M/c$ system in steady state, the PGF of the number in the queue is given by

$$G(z) = \frac{1 - B^*[\lambda(1-z)/c]}{B^*[\lambda(1-z)/c] - z} p_{c-1} + \sum_{i=0}^{c-1} p_i,$$

where $B^*(s) = \mu/(\mu + s)$ is the LST of service-time distribution.

3.9. $M/M/c$ queue with servers' vacations (Levy and Yechiali, 1976).
Consider an $M/M/c$ system in which a server proceeds on vacation when he has no unit to serve (the length of time he is on vacation being given by an exponential random variable with parameter θ) and in which the server proceeds on another vacation if he finds the queue empty on return. Suppose that the system is in steady state.
Find the joint distribution of the number of busy servers β and the number of customers N in the system

$$Pr(N = k, \beta = r), \quad r = 0, 1, 2, \ldots, \quad k \geq r.$$

Show that the average number of busy servers is λ/μ (which is the same as the average number of busy servers in an $M/M/c$ queue).
Show that the number of customers N in the system when all servers are on vacation has a geometric distribution given by

$$Pr(N = k) = \frac{c\theta}{\lambda + c\theta} \left(\frac{\lambda}{\lambda + c\theta} \right)^k.$$

Show that for $c = 2$, the average number of customers in the system is given by

$$L = \frac{\rho}{1 - \rho} + \frac{\alpha\lambda[\lambda(1 - z_1) + \theta]}{\theta}, \quad \rho = \frac{\lambda}{2\mu},$$

where

$$\alpha = \left[\lambda\left(1 - z_1^2\right) + 2\theta z_1\right]^{-1} \quad \text{and}$$

$$z_1 = \frac{(\lambda + \mu + \theta) - \{(\lambda + \mu + \theta)^2 - 4\lambda\mu\}^{1/2}}{2\lambda}.$$

(For queues with vacation, see Section 8.3.)

3.10. $M/M/1$ queue: waiting time in the system for an arrival at instant t, (virtual waiting time in the system).

Show that

$$F(x,t) = P\{W_s \le x \mid t\}$$
$$= 1 - e^{-\mu x} \sum_{n=0}^{\infty} \sum_{s=0}^{n} \frac{(\mu x)^s}{s!} p_n(t),$$

and the PDF of W_s is given by

$$w(x,t) = \frac{d}{dx} F(x,t)$$
$$= e^{-\mu x} \sum_{n=0}^{\infty} \sum_{s=0}^{n} \frac{(\mu x)^s}{s!} p_n(t) - e^{-\mu x} \sum_{n=0}^{\infty} \sum_{s=1}^{n} \frac{(\mu x)^{s-1}}{(s-1)!} p_n(t).$$

3.11. Suppose that a machine breaks down, independently of others, in accordance with a Poisson process, the average length of time for which a machine remains in working order being 36 hours. The duration of time required to repair a machine has an exponential distribution with mean 1 hour. Suppose that there are 10 machines and 1 mechanic. Find

(i) the probability that five or more machines will remain out of order at the same time;

(ii) the average number of machines in working order;

(iii) the fraction of time, on the average, the mechanic will be busy; and

(iv) the average duration of time for which a machine is not in working order.

3.12. Consider a machine repair problem with c repairpeople, m machines ($c < m$), and exponential working time and repair time having rates λ and μ, respectively. Suppose that m is very large. Show that (with notations as in Section 3.8.2)

(i) All the $a_0, a_1, \ldots, a_{c-1}$ approach zero; so also do $p_0, p_1, \ldots, p_c$ (so that all the c repairpeople are almost 100% busy at all times).

(ii) The distribution of the number of machines in working order is approximately given by the distribution of the number of busy servers in an $M/M/\infty$ queue with mean interarrival time $1/c\mu$ and mean service time $1/\lambda$.

(iii) The LST $w^*(s)$ approaches

$$\left[\exp\left(\frac{s}{\lambda}\right)\right]\left[\left(\frac{\mu}{s+c\mu}\right)\right]\left[\frac{c\mu}{s+c\mu}\right]^{m-c}$$

and the response time is asymptotically and approximately normal with

$$\text{mean} = \frac{m}{c\mu} - \frac{1}{\lambda} \quad \text{and}$$

$$\text{variance} = \frac{m-c}{(c\mu)^2} + \frac{1}{\mu^2} \quad \text{(Wong, 1979).}$$

3.13. Transient solution of an $M/M/1$ queue; alternative approach (Parthasarathy, 1987). Define

$$q_k(t) = \{\exp(\lambda+\mu)t\}[\mu p_k(t) - p_{k-1}(t)], \quad k = 1, 2, \ldots$$
$$= 0, \quad k = 0, -1, -2 \ldots$$
$$\alpha = 2\sqrt{\lambda\mu}, \qquad \beta = \sqrt{\frac{\lambda}{\mu}} = \sqrt{\rho}$$

$I_n(t)$ is a modified Bessel function of order n. Assume that $Pr\{N(0) = a\} = 1$. Show that

$$q_k(t) = \mu\beta^{k-a}(1-\delta_{0a})[I_{n-a}(\alpha t) - I_{n+a}(\alpha t)]$$
$$+ \lambda\beta^{k-a-1}[I_{n+a+1}(\alpha t) - I_{n-a-1}(\alpha t)]$$

and that for, $n = 1, 2, \ldots,$

$$p_n(t) = Pr(\text{queue length at time } t \text{ is } n)$$
$$= \left(\frac{1}{\mu}\right)\exp\{-(\lambda+\mu)t\}\sum_{k=1}^{n} q_k(t)\rho^{n-k} + \rho^n p_o(t) \quad \text{and}$$

$$p_o(t) = \int_0^t q_1(y)\exp\{-(\lambda+\mu)y\}dy + \delta_{0a}.$$

Note: See also Syski (1988). The above solution involves one integral and one finite series. The result is shown (by Syski) to be equivalent to Cohen's result (1982, (4.31), p. 82) involving three integrals.

3.14. $M/M/1$: Two-dimensional state model

Let the state of the system by time t be given by the ordered pair (i, j), where i is the number of arrivals and j is the number of departures by time t, and let $p_{ij}(t)$ denote the probability that the system is in state (i, j) by time t. Let $f_{ij}(s)$ be the LT of $p_{ij}(t)$.

Show that $p_{ij}(t)$ satisfy the differential equations

$$p'_{00}(t) = -\lambda p_{00}(t) \tag{A}$$

$$p'_{i0}(t) = \lambda p_{i-1,0}(t) - (\lambda+\mu)p_{i0}(t), \quad i \geq 1, \tag{B}$$

$$p'_{ii}(t) = \mu p_{i,i-1}(t) - \lambda p_{ii}(t), \quad i \geq 1, \tag{C}$$

$$p'_{ij}(t) = \mu p_{i,j-1}(t) + \lambda p_{i-1,j}(t) - (\lambda+\mu)p_{ij}(t), \tag{D}$$
$$i \geq 2, \quad 1 \leq j < i.$$

Show by induction that $f_{ij}(s)$ is given by

$$f_{0,0}(s) = \frac{1}{\lambda + s}$$

$$f_{ij}(s) = \left(\frac{\lambda}{\lambda + \mu + s}\right)^i \left(\frac{\mu}{\mu + s}\right)^j \sum_{k=0}^{j} \frac{(i-k)(i+k-1)!}{k!i!} \times \frac{(\lambda + s)^{k+1}}{(\lambda + \mu + s)^k}, \qquad \text{for } i \geq 1,\ 0 \leq j \leq i. \quad \text{(E)}$$

Show further that

$$p_{0,0}(t) = e^{-\lambda t}$$

and for $i \geq 1, 0 \leq j \leq i$,

$$p_{ij}(t) = \left(\frac{\lambda}{\mu}\right)^i \frac{(\mu t)^j e^{-\lambda t}}{i!} \sum_{k=0}^{j} \frac{(i-k)}{k!} \times \left\{ \sum_{m=0}^{j-k} \frac{(-1)^m (m+i+k-1)!}{m!(j-k-m)!(\mu t)^{m+k}} \left[1 - e^{-\mu t} \sum_{r=0}^{m+i+k-1} \frac{(\mu t)^r}{r!} \right] \right\}$$

$$= \frac{(\lambda t)^i (\mu t)^j \exp(-\lambda t - \mu t)}{i!\, j!} \left[\frac{j}{\mu t} + \left(1 - \frac{j}{\mu t}\right) \frac{i!}{(\mu t)^i} \sum_{r=i}^{\infty} \frac{(\mu t)^r}{r!} \right]. \quad \text{(F)}$$

Again $p_n(t) = P\{N(t) = n\}$ can be obtained from $p_{i,j}(t)$ as follows.

$$P_n(t) = P\{N(t) = n\} = \sum_{j=0}^{\infty} p_{n+j,j}(t). \quad \text{(G)}$$

Show further that the fraction of time the server is idle until time t is given by

$$I(t) = \frac{1}{t} \int_0^t \sum_{j=0}^{\infty} p_{jj}(\tau) d\tau$$

and the fraction of time the server is busy until time t is given by

$$B(t) = 1 - I(t)$$

(Pegden and Rosenshine, 1982; Hubbard *et al.*, 1986). See also Boxma (1984).

3.15. The transient state distribution of an $M/M/1$ queue has been obtained in a different form (by Sharma) as given below.
Show that the probability $p(n, t)$ that there are n (customers) in the system $(M/M/1)$ at time t, given that the system was empty at time $t = 0$, is given by

$$p(n,t) = (1-\rho)\rho^n + \rho^n \exp\{-(\lambda+\mu)t\} \times \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \left\{ \sum_{m=0}^{n+k} (k-m) \frac{(\mu t)^m}{m!} \right\}$$

(the corresponding result, given that the system starts with i customers at time $t = 0$, is also given).

From the above, as $t \to \infty$ (when $\rho < 1$), the steady-state result follows.

Sharma and Gupta (1982) consider also the transient behavior of $M/M/1/K$ queue with finite buffer and obtain $p(n, t), 0 \le n \le k$, as the sum of two terms as above. (See Sharma (1997).)

3.16. Transient output distribution for an $M/M/c$ system.
Let $R_t(k \mid n, m)$ etc., be defined as in Section 3.10.3, and let

$$R^*(k \mid n, m) = R_z^*(k \mid n, m) = \int_0^\infty e^{-zt} R_t(k \mid n, m)\, dt$$

be the LT of $R_t(k \mid n, m)$. Then (3.10.40) can be put as

$$[z + \mu \min(c, m+n-k)]\, R^*(k \mid n, m) = R^*(k \mid n-1, m) + \mu \min(c, m+n-k-1)\, R^*(k-1 \mid n, m). \quad \text{(A)}$$

For the initial condition $P_t(0 \mid 0, m) = \exp\{-\mu \min(c, m)t\}$ show that the solution of (A) is given by

$$R^*(k \mid n, m) = \left[\frac{\prod_{j=n-k+1} \{z + \mu \min(c, m+j)\}}{\prod_{j=0}^{n-k} \{z + \mu \min(c, m+j)\}} \right] \times \sum_{i_k=0}^{n} \sum_{i_{k-1}=0}^{i_k} \cdots \sum_{i_k=0}^{i_2} W(m-k+1+i_k) \quad \text{(B)}$$

$$\times W(m-k+2+i_{k-1}) \cdots W(m-1+i_2)\, W(m+i_1), \quad \text{(B1)}$$

where $\prod_{j=p}^{q}(.)$ is defined to be 1 whenever $q < p$,
and

$$W(i) = \begin{cases} \dfrac{\mu \min(c,i)}{[z+\mu \min(c,i-1)][z+\mu \min(c,i)]}, & i \ge 1 \\ 0, & i \le 0. \end{cases}$$

Further, show that

$$\lim_{t\to\infty}(\mu t)^{n-k}P_t(k\mid n,0)=\lim_{t\to\infty}\left(\frac{n!\mu^{n-k}}{t^k}\right)R_t(k\mid n,0)$$

$$=\frac{n!\mu^{n-k}}{k!}\lim_{z\to 0}z^{k+1}R_z^*(k\mid n,0). \qquad \text{(C)}$$

An explicit solution of $P_t(k \mid n, 0)$ when $N(0) = m = 0$ is also given by Everitt and Downs (1984).

3.17. Multiserver queue with balking and reneging.
Consider a c-server queueing system with Poisson input with rate λ and exponential service time with rate μ for each of all c-servers. Suppose that (i) an arriving customer who finds all the c-servers busy on arrival may balk (leave without joining the system) with probability q or may join the system with probability $p(=1-q)$; (ii) after joining the queue, a customer may renege independently of others; he waits for a random length of time for service to begin, the length of time being an exponential random variable with parameter α; otherwise he departs; and (iii) a customer who balks or reneges and decides to return later is considered as a new arrival independent of his previous balking or reneging.

Find the differential-difference equations of the state $N(t)$ of the system. If $p_n(t) = P\{N(t) = n\}$ and $\lim_{t\to\infty} p_n(t) = p_n$, then show that

$$p_n=\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, \quad n\le c,$$

$$=\frac{(\lambda p)^{n-c}}{(c\mu+\alpha)(c\mu+2\alpha)\ldots[c\mu+(n-c)\alpha]}p_c, \quad n>c.$$

Find p_n when there is only balking and no reneging. Also deduce Erlang's loss formula (Haghighi-Montazer *et al.*, 1986). See also Abou-El-Ata and Hariri (1992) and Mohanty *et al.* (1993) for some generalizations.

3.18. Two-channel model with ordered entry with $M = N = 1$.
Suppose that the service rates at the two channels are different, being μ_1 and μ_2 at the first and second channel, respectively. Show that the steady-state probabilities are given by

$$p_{0,1}=\frac{\lambda^2\mu_1}{C}$$

$$p_{1,0}=\frac{\lambda\mu_2(\lambda+\mu_1+\mu_2)}{C},$$

$$p_{1,1}=\frac{\lambda^2(\lambda+\mu_2)}{C}, \quad \text{and}$$

$$p_{0,0}=\frac{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)}{C},$$

where $C = (\lambda + \mu_1)[(\lambda + \mu_2)^2 + \mu_1\mu_2]$. Show further that the waiting time W in steady state has the distribution given by

$$P\{W \le t\} = [\{\mu_1\mu_2(2\lambda + \mu_1 + \mu_2) + \lambda^2\mu_1\}\{1 - e^{-\mu_1 t}\} + \lambda\mu_2(\lambda + \mu_1 + \mu_2)(1 - e^{-\mu_2 t})]/C, \quad t > 0$$

(Disney, 1962). Obtain the corresponding results for the particular case $\mu_1 = \mu_2$.

3.19. Two-channel model with ordered entry and with $M = 1, N = 3$.
Let $P_{i,j}$ denote the probability that the number in channel 1 is i and that the number in channel 2 is j (i, j include those in service, if any).
Show that

$$P_{0,0} = \frac{4(\rho + 1)(\rho + 2)}{A} \qquad P_{1,0} = \frac{\rho(\rho^2 + 8\rho + 8)}{A}$$

$$P_{0,1} = \frac{\rho^2(3\rho + 4)}{A} \qquad P_{1,1} = \frac{\rho^2(\rho + 1)(\rho + 4)}{A}$$

$$P_{0,2} = \frac{2\rho^3(\rho + 1)}{A} \qquad P_{1,2} = \frac{\rho^3(\rho + 1)(\rho + 2)}{A}$$

$$P_{0,3} = \frac{\rho^4(\rho + 1)}{A} \qquad P_{1,3} = \frac{\rho^4(\rho + 1)^2}{A},$$

where

$$A = \rho^6 + 4\rho^5 + 8\rho^4 + 13\rho^3 + 20\rho^2 + 20\rho + 8, \quad \rho = \frac{\lambda}{\mu}$$

(Disney, 1962).

3.20. $M/M/c$ queue ($c \ge 1$): transient state distribution.
Suppose that $k(\ge 1)$ customers are already present at time $t_0 = 0$ and that the nth new customer arrives at time t_n ($n \ge 1$).
Denote

$$X_n = \text{number of customers present at time } t_n + 0 \text{ (including the } n\text{th arrival at } t_n).$$

$$P_k(n, i) = Pr\{X_n = i \mid \text{number present at time } t_0(=0) \text{ is } k\}$$

$$\rho = \lambda/c\mu \text{ (which need not be } <1).$$

(The first arriving customer at t_1 need not find k customers present but will find j ($0 \le j \le k$) customers.)

If $k \geq 1$ and $n \geq 1$, show that

$$P_k(n, k+n) = \left[\frac{\rho}{(\rho+1)}\right]^n \quad \text{if } k \geq c$$

$$= \frac{\rho^n}{\prod_{j=1}^{n}\left[\rho + \frac{k+j-i}{c}\right]} \quad \text{if } k+n \leq c$$

$$= \frac{\rho^n}{\left[(\rho+1)^{n-c+k}\prod_{j=1}^{c-k}\left\{\rho + \frac{k+j-i}{c}\right\}\right]} \quad \text{if } k < c < k+n.$$

If $k = 0$, then for $n \geq 1$, show that

$$P_0(n, n) = \frac{\rho^n}{\prod_{j=1}^{n}\left[\rho + \frac{j-1}{c}\right]} \quad \text{if } n \leq c$$

$$= \frac{\rho^n}{\left[(\rho+1)^{n-c}\prod_{j=1}^{c}\left\{\rho + \frac{(j-1)}{c}\right\}\right]} \quad \text{if } n > c.$$

If $k \geq 1$, then for $2 \leq i \leq k$

$$P_k(1, i) = \left[\frac{\rho}{\left\{\rho + \frac{(i-1)}{c}\right\}}\prod_{j=1}^{k-i+1}\left\{1 - \frac{\rho}{\left\{\rho + \frac{(k-j+1)}{c}\right\}}\right\}\right] \quad \text{if } k \leq c$$

$$= \frac{\rho}{(\rho+1)^{k-i+2}}, \quad k > c, \ i > c.$$

$$= \left[\frac{\rho}{\left\{(\rho+1)^{k-c+1}\left[\frac{\rho+(i-1)}{c}\right]\right\}}\right]\left[\prod_{j=1}^{c-1}\left\{1 - \frac{\rho}{[\rho + (c-j)/c]}\right\}\right],$$

$$i \leq c \leq k.$$

Let D_n = waiting time (or delay) in queue of the nth customer, and $G_q(x; n)$ = DF of an Erlang-q RV with mean q/η

$$= 1 - \{\exp(-\eta x)\}\sum_{j=0}^{q-1}\frac{(\eta x)^j}{j!}, \quad x \geq 0.$$

Then show that

$$F_n(x) = P(D_n \leq x) = \sum_{i=1}^{s} P_i(n, i) + \sum_{i=c+1}^{k+n} G_{i-c}(x; c\mu)\, P_k(n, i), \quad \text{and}$$

$$E(D_n) = \left(\frac{1}{c\mu}\right)\left[\sum_{i=c+1}^{k+n}(i-c)\, P_k(n, i)\right].$$

(See Kelton and Law (1985), who also consider the implication for steady-state simulation.)

3.21. Multichannel queue with ordered entry and heterogeneous servers (Mastsui and Fukuta, 1977).

Consider an ordered entry Poisson input queue (with rate λ) with c exponential servers, the ith server having rate μ_i. Suppose that the system is in steady state and that there is no waiting line before any of the servers. Denote

$$m_i = \frac{\mu_i}{\lambda}, \quad i = 1, 2, \ldots, c$$

P_0 = probability that the system is idle

$P_{i_1,\ldots,i_k}$ = probability that i_1th, $\ldots$, i_kth channels are busy and others are idle ($i_k < c$)

$P_{1,2,\ldots,c}$ = probability that the system is completely busy (with all the servers busy).

This gives the overflow probability.

Show that for $c = 2$

$$P_{1,2} = \frac{1 + m_2}{(1 + m_1)\{(1 + m_2)^2 + m_1 m_2\}}$$

and that the faster server should be assigned to the first channel to decrease the overflow probability (as should be intuitively clear).

Examine the case $c = 3$.

Note: See Yao (1987) for further results of such a system; also for comparison of various server arrangements and development of partial order.

3.22. Consider a three-channel Poisson queue with ordered entry having no waiting space (as in Section 3.11.3). Find the steady-state probabilities and verify the results with those of the corresponding loss system.

3.23. Multiserver Poisson queue with ordered entry (Nawijn, 1983). Consider the following two c-channel systems with ordered entry such that the c-channels are numbered $1, 2, \ldots, c$ and an arriving customer who finds a free channel joins the one with the lowest index:

(A) $M/M/c$ system (*queueing* or *delay system*)

(B) $M/M/c/c$ system (*loss system*)

Denote

$p = \lambda/\mu, \quad \rho = p/c < 1$

N = number in the $M/M/c$ system (A)

$B(k, p)$ = Erlang's loss formula for $M/M/k/k$, $k = 1, 2, \ldots, c$ (see Eq. 3.7.3)

u_k = utilization factor of channel k in system (A)

v_k = utilization factor of channel k in system (B)

Show that, for $k = 1, 2, \ldots, c$,

$$v_k = p[B(k-1, p) - B(k, p)], \quad \text{and}$$

$$u_k = v_k \, Pr\{N \leq c\} + Pr\{N > c\}.$$

Verify that $\{v_k\}$ is monotone decreasing in k. Give an intuitive explanation. Find $\sum_{k=1}^{c} u_k$ and interpret the result.

3.24. Loss formula for $M/M/c/c + r, r \geq 0$ Model.
This has been discussed by Pacheco (1994b). It can be shown that the loss formula for integral c and r becomes

$$B(c, a; r) = \frac{(a^c/c!)(a/c)^r}{\sum_{k=0}^{c} a^k/k! + (a^c/c!)\sum_{i=1}^{r}(a/c)^i}, \quad a = \lambda/\mu.$$

It may be noted that the second term in the denominator vanishes when $r = 0$; one then gets the Erlang Loss Formula.

3.25. Transient behavior of $M/M/c/c$ Loss Model.
This topic has received attention of late. Abate and Whitt (1998) discuss it by using numerical transform inversion.
The rate of convergence of the $M/M/c/c$ model has been studied in a recent paper by Fricker *et al.* (1999) using martingales and coupling techniques.
It can be shown that if $N_c(t)$ is the number of customers in the system at time t and if

$$\bar{N}_c(t) = N_c(t)/c$$

$$\bar{N}(0) = \lim_{c \to \infty} \frac{N_c(0)}{c},$$

then

$$\bar{N}_c(t) = \min\{a + (\bar{N}(0) - a)e^{-t}, 1\}.$$

It follows that

(1) when $a > 1$, the queue becomes full after a finite time;

(2) when $a < 1$, the queue is never full; and

(3) when $a = 1$ ($\lambda = \mu$), the queue becomes full at infinity.

How do you interpret these in terms of Erlang loss probability?

3.26. $M/M/c$ queue with impatient customers

Suppose that in an $M/M/c$ queue, each arriving customer enters the system but is only willing to wait in queue for a fixed time $T > 0$, after which the customer leaves the system and is lost (if his service has not begun after his waiting for time T).

Show that the probability of loss is given by

$$P_{loss}(T) = \frac{(1-\rho)(1-W_q(T))}{1-\rho(1-W_q(T))},$$

where $W_q(.)$ is the DF of waiting (queueing) time in the standard $M/M/c$ queue. (See Boots and Tijms (1999).)

Results for general service time are given in Bocharov and Pechinkin (1995).

References and Further Reading

Abate, J., and Whitt, W. (1987). Transient behavior of $M/M/1$ queue starting at the origin. *Queueing Systems* **2**, 42–66.

Abate, J., and Whitt, W. (1988). Transient behavior of the $M/M/1$ queue via Laplace transforms. *Adv. Appl. Prob.* **20**, 145–178.

Abate, J., and Whitt, W. (1998). Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Comm. Statist. Stoch. Models* **14**, 663–680.

Abou-El-Ata, M. O., and Hariri, A. M. A. (1992). The $M/M/c/N$ queue with balking and reneging. *Computers Ops. Res.* **19**, 713–716.

Abramowitz, M., and Stegun, I. A. (Eds.) (1965). *Handbook of Mathematical Functions*, Dover Publications, New York.

Albin, S. L. (1984). Analyzing $M/M/1$ queues with perturbations in the arrival process. *J. Opnl. Res. Soc.* **35**, 303–309.

Bacceli, F., and Massey, W. A. (1989). A sample path analysis of the $M/M/1$ queue. *J. Appl. Prob.* **26**, 418–422.

Bailey, N. T. J. (1954). A continuous time treatment of a simple queue using generating functions. *J.R.S.S.* **B16**, 288–291.

Baker, K. R. (1973). A note on the operating policies for the $M/M/1$ queue with exponential startups. *INFOR* **11**, 71–72.

Berezner, S. A., Krzensinski, A. A., and Taylor, P. G. (1998). On the inverse of Erlang function. *J. Appl. Prob.* **35**, 246–252.

Bocharov, P. P., and Pechinkin, A. V. (1995). *Theory of Queues*, Peoples' Friendship Univ. of Russia Publ., Moscow.

Böhm, W., and Mohanty, S. G. (1990). On the transient solution of N-policy queues. *Statistics Research Report*, No. 11, McMaster University, Hamilton, Ontario.

Boots, N. K, and Tijms, H. C. (1999). A multiserver queueing system with impatient customers. *Manag. Sc.* **45**, 444–448.

Borthakur, A., Medhi, J., and Gohain, R. (1987). Poisson input queueing system with startup time and under control operating policy. *Comp. Opns. Res.* **14**, 33–40.

Boxma, O. J. (1984). The joint arrival and departure process for the $M/M/1$ queue. *Stat. Neerlandica* **38**, 199–208.

Bunday, B. D., and Scarton, R. E. (1980). The $G/M/r$ machine interference model. *Euro. J. Opnl. Res.* **4**, 399–402.

Burke, P. J. (1956). Output of a queueing system. *Opns. Res.* **4**, 699–704.
Burke, P. J. (1964). The dependence of delays in tandem queues. *Ann. Math. Stat.* **35**, 874–875.
Burke, P. J. (1968). The output process of a stationary $M/M/s$ queueing system. *Ann. Math. Stat.* **39**, 1144–1152.
Burke, P. J. (1972). Output processes and tandem queues in *Computer Communication Networks and Teletraffic* (Ed. J. Fox) Polytechnic Press, New York.
Champernowne, D. G. (1956). An elementary method of solution of the queueing problem with a single server and constant parameters. *J.R.S.S.* **B18**, 125–128.
Chaudhry, M. L., and Templeton, J. G. C. (1973). A note on the distribution of a busy period for $M/M/c$ queueing system. *Math Oper U. Stat.* **1**, 75–79.
Cohen, J. W. (1982). *The Single Server Queue*, 2nd ed., North-Holland, Amsterdam, The Netherlands.
Conolly, B. W. (1958). A difference equation technique applied to the simple queue with arbitrary arrival interval distribution. *J.R.S.S.* **B21**, 168–175.
Conolly, B. W., and Langaris, C. (1993). On a new formula for the transient state probability for $M/M/1$ queue and computational implications. *J. Appl. Prob.* **30**, 237–246.
Cooper, R. B. (1976). Queues with ordered servers that work at different rates. *Opsearch* **13**, 69–78.
Cooper, R. B. (1981). *Introduction to Queueing Theory*, 2nd ed., North Holland, Amsterdam.
Cooper, R. B. (1990). Queueing Theory, in *Handbooks in Operations Research and Management Science*, vol. 2, pp. 469–518 (Eds. D. P. Heyman and M. J. Sobel), North Holland, Amesterdam.
Cooper, R. B., and Solomon, M. K. (1984). The average time until bucket overflow. *ACM Trans. Database Syst.* **9**, 392–398.
Descloux, A. (1962). *Delay Tables for Finite and Infinite Source Systems*, McGraw Hill, New York.
De Vany, A. (1976). Uncertainty, waiting time and capacity utilization: a stochastic theory of product quality. *J. Pol. Eco.* **84**, 523–541.
Dietrich, G., Krush, W., Michel, G., Ondra, F., Peter, E., and Wanger, G. (1966). *Teletraffic Engineering Manual*, Standard Electrik, Lorenz, Stutgart, Federal Republic of Germany.
Disney, R. L. (1962). Some multichannel queueing problems with ordered entry. *J. Industrial Eng.* **13**, 46–48.
Disney, R. L. (1963). Some multichannel queueing problems with ordered entry. An application to conveyor theory. *J. Industrial Eng.* **14**, 105–108.
Disney, R. L., and König, D. (1985). Queueing networks: a survey of their random processes. *SIAM Review* **27**, 335–403.
Dufkova, V., and Zitek, F. (1975). On a class of queue disciplines. *Aplikace Matmatiky Sv.* **20**, 345–358.
Elsayed, E. A. (1983). Multichannel queueing systems with ordered entry and finite source. *Comp. & Opns. Res.* **10**, 213–222.
Elsayed, E. A., and Elayat, H. A. (1976). Analysis of closed-loop conveyor systems with multiple Poisson inputs and outputs. *Int. J. Prod. Res.* **14**, 99–107.
Elsayed, E. A., and Proctor, C. L. (1977). Ordered entry and random choice conveyors with multiple Poisson input. *Int. J. Prod. Res.* **15**, 439–451.
Everitt, D. E., and Downs, T. (1984). The output of the $M/M/s$ queue. *Opns. Res.* **32**, 796–808.
Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, vol. 2, Wiley, New York.
Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed., Wiley, New York.
Flatto, L. (1997). The waiting time distribution for the random order service $M/M/1$ queue. *Ann. Appl. Prob.* **7**, 382–409.
Fortet, R. (1948). Sur la probabilité de perte d'un appel téléphonique. *C. R. Acad. Sc. Paris* **226**, 1502–1504.

Fricker, C., Phillipe, R., and Danielle, T. (1999). On the rates of convergence of Erlang model. *J. Appl. Prob.* **36**, 1167–1184.

Goodman, J. B., and Massey, W. A. (1984). The non-ergodic Jackson network. *J. Appl. Prob.* **21**, 266–277.

Grassman, W. (1983). The convexity of the mean queue size of the $M/M/c$ queue with respect to the traffic intensity. *J. Appl. Prob.* **20**, 252–267.

Gray, J. P., and Mcneill, T. B. (1979). SNA multiple system networking. *IBM Syst. J.* **18**, 263–297.

Greenberg, H., and Greenberg, I. (1966). The number served in a queue. *Opns. Res.* **14**, 137–144.

Gregory, G., and Litton, C. D. (1975). A conveyor model with exponential service times. *Int. J. Prod. Res.* **13**, 1–7.

Gullemin, F., and Simonian, A. (1995). Transient characteristics of $M/M/\infty$ system. *Adv. Appl. Prob.* **27**, 862–888.

Gupta, S. K. (1966). Analysis of two-channel queueing system with ordered entry. *J. Industrial Engineering* **17**, 54–55.

Haghighi-Montazer, A., Medhi, J., and Mohanty, S. G. (1986). On a multi-server Markovian queueing system with balking and reneging. *Comp. Opns. Res.* **13**, 421–425.

Haight, F. A. (1961). A distribution analogous to the Borel-Tanner. *Biometrika* **48**, 167–173.

Harel, A. (1987). Sharp bounds and simple approximations for the Erlang delay and loss formulas. *Mgmt. Sci.*

Harel, A., and Zipkin, P. (1987). Strong convexity results for queueing systems. *Opns. Res.* **35**, 405–418.

Harris, R. (1974). The expected number of idle servers in a queueing system. *Opns. Res.* **22**, 1258–1259.

Hubbard, J. R., Pegden, C. D., and Rosenshine, M. (1986). The departure process of an $M/M/1$ queue. *J. Appl. Prob.* **23**, 249–255.

Hunt, P., and Kurtz, T. (1994). Large loss networks. *Stoch. Prob. Appl.* **53**, 363–378.

Jackson, R. R. P., and Henderson, J. C. (1966). The time-dependent solution to the many server Poisson queue. *Opns. Res.* **14**, 720–723.

Jagerman, D. L. (1974). Some properties of the Erlang loss function. *Bell Syst. Tech. J.* **53**, 525–551.

Jagerman, D. L., Melamed, B., and Willinger, W. (1997). Stochastic Modeling for Traffic Processes in *Frontiers in Queueing* (Ed. J. H. Dshalalow), 271–320, CRC Press, Boca Raton, FL.

Jagers, A. A., and Van Doorn, E. A. (1986). On the continued Erlang function. *Opers. Res. Lett.* **5**, 43–46.

Jordan, C. (1950). *The Calculus of Finite Differences*, Chelsea, New York.

Kalashnikov, V. V., and Rachev, S. T. (1990). *Mathematical Methods for Construction of Queueing Models*, Wadsworth & Brooks, Belmont, CA.

Kaufman, J. S. (1979). The busy probability in $M/G/N/N$ loss systems. *Opns. Res.* **27**, 204–210.

Kelly, F. (1991). Loss networks. *Ann. Appl. Prob.* **1**, 319–378.

Kelton, W. D., and Law, A. M. (1985). The transient behavior of $M/M/s$ queue with implications for steady state simulation. *Opns. Res.* **33**, 378–396.

Khinchin, A. Y. (1960). *Mathematical Models in the Theory of Queueing*, Griffin, London.

Kijima, M., Abate J., and Whitt, W. (1991). Decompositions of the $M/M/1$ transient function, *QUESTA* **9**, 323–336.

Kingman, J. F. C. (1962). On queues in which customers are served in random order, *Proc. Camb. Phil. Soc.* **58**, 79–91.

Kleinrock, L. (1975). *Queueing Systems*, vol. 1, Wiley, New York.

Knessl, C. (1990). On the transient behaviour of the $M/M/m/m$ loss model. *Comm. Statist. Stoch. Models* **6**, 749–776.

Kobayashi, H., and Mark, B. L. (1997). Product Form Loss Networks in *Frontiers in Queueing* (Ed. J. H. Dshalalow), 147–195, CRC Press, Boca Raton, FL.

Koenigsberg, E. (1980). Cycle queue models of semi-conductor noise and vehicle fleet operations. *Int. J. Electr.* **48**, 83–91.

Krishnan, K. R. (1990). The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates, *IEEE Trans. Comm.* 38#9.

Laguesdron, P., Pellaumail, J., Rubino, G., and Sericola, B. (1993). Transient analysis of the $M/M/1$ queue. *Adv. Appl. Prob.* **25**, 702–713.

Ledermann, W., and Reuter, G. E. H. (1954). Spectral theory for the differential equations of simple birth and death processes. *Phil. Trans.* **A246**, 321–369.

Lee, H. (1990). A short note on the Poisson input queueing system with start-up time and under control-operating policy. *Computers Ops. Res.* **17**, 119–121.

Levy, Y., and Yechiali, U. (1976). An $M/M/s$ queue with servers vacations. *INFOR* **14**, 153–163.

Lin, B. W., and Elsayed, E. A. (1978). A general solution for multichannel queueing systems with ordered entry. *Int. J. Comp. Opns. Res.* **5**, 219–225.

Martins-Neto, A. F., and Wong, E. (1976). A Martingale Approach to Queues in *Stochastic Systems: Modeling, Identification and Optimization* (Ed. R. J.-P. Wets), North-Holland, Amsterdam, The Netherlands.

Matsui, M., and Fukuta, J. (1977). On a multichannel queueing system with ordered entry and heterogeneous servers. *AIIE Trans.* **9**, 209–214.

Medhi, J. (1994). *Stochastic Processes*, 2nd ed., Wiley, New York and Wiley Eastern (Now, New Age Int. Publishers), New Delhi.

Medhi, J. (2002). The evergreen Erlang loss function (under submission).

Messerli, E. J. (1972). Proof of a convexity property of the Erlang's B-formula. *Bell Syst. Tech. J.* **51**, 951–953.

Mohanty, S. G., and Jain, J. L. (1971). The distribution of the maximum queue length, the number of customers and the duration of the busy period for the queueing system $M/M/1$ involving batches. *INFOR* **9**, 161–166.

Mohanty, S. G., and Panny, W. (1990). A discrete time analogue of the $M/M/1$ queue and the transient solution, I & II. *Statistics Research Report* No. 9, Mcmaster University, Hamilton, Ontario.

Mohanty, S. G., Montazer-Haghighi, A., and Trueblood, R. (1993). On the transient behavior of a finite birth–death process with an application. *Computers Ops. Res.* **20**, 239–248.

Morisaku, T. (1976). Techniques for date-truncation in digital computer simulation. Ph.D. dissertation, U. of Southern California.

Morse, P. M. (1955). Stochastic properties of waiting lines. *Opns. Res.* **3**, 255–261.

Muth, E. J., and White, J. A. (1979). Conveyor theory: A survey. *AIIE Trans.* **11**, 270–277.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica* **37**, 15–23.

Nawijn, W. M. (1983). A note on many-server queueing systems with ordered entry, with an application to conveyor theory. *J. Appl. Prob.* **20**, 144–152.

Nawijn, W. M. (1984). On a two-server finite queueing system with ordered entry and deterministic arrivals. *Euro. J. Opnl. Res.* **18**, 388–395.

Neuts, M. F. (1974). The Markov Renewal Branching Processes in *Math, Methods in Queueing Theory* (Ed. A. B. Clarke), Lecture Notes in Ec. & Math. Systems, No. 98, Springer-Verlag, New York.

Neuts, M. F. (1976). Moment formulas for Markov renewal branching processes. *Adv. Appl. Prob.* **8**, 690–711.

Neuts, M. F. (1979). Queues solvable without Rouché's theorem. *Opns. Res.* **27**, 767–781.

Newell, G. F. (1984). *The $M/M/\infty$ Service System with Ranked Servers in Heavy Traffic*, Springer-Verlag, New York.

O'Brien, G. G. (1954). Some queueing problems. *J. Soc. Ind. Appl. Math.* **2**, 134.

Pacheco, A. (1994a). Second order properties of the loss probability in $M/M/s/s+c$ systems. *Queueing Systems* **15**, 289–308.

Pacheco, A. (1994b). Some properties of the delay probabilities in $M/M/s/s + c$ systems. *Queueing Systems* **15**, 309–324.

Palm, C. (1943). Intensitatsschwankungen im Fernsprechverkehr. *Ericsson Technics* **44**, 1–189.

Parthasarathy, P. R. (1987). A transient solution to an $M/M/1$ queue: a simple approach. *Adv. Appl. Prob.* **19**, 997–998.

Parthasarathy, P. R., and Sharafali, M. (1989). Transient solution to the many server Poisson queue: a simple approach. *J. Appl. Prob.* **26**, 584–594.

Pegden, C. D., and Rosenshine, M. (1982). Some new results for the $M/M/1$ queue. *Mgmt. Sci.* **28**, 821–828.

Pourbabai, B. (1987). Approximation of the overflow process from a $G/M/N/K$ queueing system. *Mgmt. Sci.* **33**, 931–938.

Pourbabai, B., and Sonderman, D. (1986). Server utilization factors in queueing loss systems with ordered entry and heterogeneous servers. *J. Appl. Prob.* **23**, 236–242.

Prabhu, N. U. (1965). *Queues and Inventories*, Wiley, New York.

Pritsker, A. A. B. (1966). Applications of multichannel queueing results to the analysis of conveyor systems. *J. Ind. Eng.* **17**, 14–21.

Proctor, C. L., Elsayed, E. A., and Elayat, H. S. (1977). A conveyor system with homogeneous and heterogeneous servers with dual input. *Int. J. Prod. Res.* **15**, 73–85.

Rachev, S. T. (1989). The problem of stability in queueing theory. *Queueing systems* **4**, 287–318.

Reich, E. (1965). Departure processes in *Proc. Symp. on Congestion Theory*, 439–457 (Eds. W. L. Smith and W. E. Wilkinson), U. of North Carolina Press, Chapel-Hill, NC.

Riordon, J. (1953). Delay curves for calls received at random, *Bell. Syst. Tech. J.* **32**, 110–119.

Rothkopf, M. H., and Rech, P. (1987). Perspective on queues. Combining queues is not always beneficial. *Opns. Res.* **35**, 906–909.

Rue, R. C., and Rosenshine, M. (1981). Optimal control for entry of many classes of customers to an $M/M/1$ queue. *Nav. Res. Log. Qrly.* **28**, 489–495.

Saaty, T. L. (1960). Time dependent solution of the many server Poisson queue. *Opns. Res.* **8**, 755–772.

Saaty, T. L. (1961). *Elements of Queueing Theory*, McGraw-Hill, New York.

Shanthikumar, J. G., and Yao, D. D. (1987). Comparing ordered entry queues with heterogeneous servers. *Queueing Systems* **2**, 235–244.

Sharma, O. P. (1997). *Markovian Queues*, Allied Publishers Ltd., New Delhi.

Sharma, O. P., and Gupta, U. C. (1982). Transient behavior of $M/M/1/N$ queue. *St. Proc. & Appl.* **13**, 327–331.

Shoraby, K. (1995). *Discrete Time Queueing Theory*, Springer-Verlag, New York.

Smith, D. R., and Whitt, W. (1981). Resource sharing for efficiency in traffic system. *Bell System Tech. J.* **60**, 39–55.

Sonderman, D. (1979). Comparing multi-server queues with finite waiting rooms. I. Same number of servers, II. Different number of servers. *Adv. Appl. Prob.* **11**, 439–447; 448–455.

Sonderman, D. (1982). An analytical model for recirculating conveyors with stochastic inputs and outputs. *Int. J. Prod. Res.* **20**, 591–605.

Sphicas, G. P., and Shimshak, D. G. (1978). Waiting time variability in some single server queueing systems. *J. Opnl. Res. Soc.* **29**, 65–70.

Syski, R. (1986). *Introduction to Congestion Theory in Telephone Systems*, 2nd ed., North-Holland, Amsterdam, The Netherlands.

Syski, R. (1988). Further comments on the solution of the $M/M/1$ queue. *Adv. Appl. Prob.* **20**, 693.

Takács, L. (1962). *Introduction to the Theory of Queues*, Oxford University Press, New York.

Takács, L. (1969). On Erlang's formula. *Ann. Math. Stat.* **40**, 71–78.

Towsley, D. (1987). An application of the reflection principle to the transient analysis of the $M/M/1$ queue. *Naval. Res. Log. Qrly.* **34**, 451–456.

Van Doorn, E. (1981). *Stochastic Monotonicity and Queueing Applications in Birth Death Processes.* Lecture Notes in Statistics, No. 4, Springer-Verlag, New York.

Vaulot, E. (1946). Delais d'attente des appels téléphoniques traités au hasard. *C. R. Acad. Sc. Paris* **222**, 268–269.

Vaulot, E. (1951). Les formules d'Erlang et leur calcul pratique. *Ann. Telecom.* **6**, 279–286.

Whitt, W. (1981). Comparing counting processes and queues. *Adv, Appl. Prob.* **13**, 207–220.

Whitt, W. (1986). Deciding which queue to join: some counterexamples. *Opns. Res.* **34**, 55–62.

Weiss, E. N., and McClain, J. O. (1987). Administrative delays in acute health care facilities: a queueing analytic approach. *Opns. Res.* **35**, 35–44.

Wong, J. W. (1979). Response time distribution of the $M/M/m/N$ queueing model. *Opns. Res.* **27**, 1196–1202.

Yao, D. D. (1986). Convexity properties of the overflow in an ordered-entry system with heterogeneous servers. *Opns. Res. Lett.* **5**, 145–147.

Yao, D. D. (1987). The arrangement of servers in an ordered-entry system. *Opns. Res.* **35**, 759–763.

Zolotarev, V. M. (1977). General problems of the stability of mathematical models. *Proc. 41st Int. Stat. Ins.*, New Delhi.

Non-Birth-and-Death Queueing Systems: Markovian Models

4.1 Introduction

In the preceding chapter, we discussed Markovian queueing processes that can be studied as birth-and-death processes. There the transitions occur to neighboring states: from state i to state $i-1$ $(i \geq 1)$ or from state i to state $i+1$ $(i \geq 0)$. Here we examine some Markovian models that arise out of non-birth-and-death processes where transitions occur from a state to a state not necessarily neighboring: from state i to, say, state $i-k$ $(i \geq k \geq 1)$ or from state i to, say, state $i+r$ $(i \geq 0, r \geq 1)$. The processes considered are Markovian, and the Chapman-Kolmogorov equations pertaining to the model can be written down and the solutions can be obtained in a similar manner. We first consider systems where, of the two distributions (the interarrival-time distribution and the service-time distribution), one is exponential, while the other is Erlangian.

4.1.1 The system $M/E_k/1$

We consider a single-channel system where the arrival process is Poisson with rate λ and the service-time distribution is E_k having density

$$b(t) = \frac{k\mu(k\mu t)^{k-1}e^{-k\mu t}}{(k-1)!}, \quad t \geq 0,$$

with mean $1/\mu$.

The service time may be thought of as consisting of k independent exponential stages, each with mean $(1/k\mu)$. As soon as a customer arrives in the system, he may be considered to have k stages of service to be completed by him.

Until he has completed all the stages of service, the next arrival will wait in the queue. The system state at an instant may be studied in terms of stages of service remaining to be completed, and the stages may be marked in the *reverse order*, the first stage corresponding to k, the second to $(k-1)$, and the last to stage 1. That is, if he is at ith stage, he has yet to complete $k-(i-1)$ stages of service.

Thus, if at any instant there are $j(>0)$ customers in the system, and the customer being served is at the ith stage $(1 \leq i \leq k)$ of service, and if r denotes the number of stages contained in the total system (or number of stages that remain to be completed) at that time, then

$$\begin{aligned} r &= (j-1)k + [k-(i-1)] \\ &= jk - i + 1, \quad r \geq 0. \end{aligned}$$

Here, a customer's arrival effects a transition from state i to state $i+k(i \geq 0)$ in the system, whereas with a service completion (departure of a customer from the system), transition occurs from state i to $i-1$ $(i \geq 1)$. The process is non-birth-and-death but may be described, following Keilson, as skip-free downward, since transitions from state i occur to a lower state only, to state $(i-1)$.

Let $N(t)$ be the number of stages in the total system at the epoch t and let

$$p_n(t) = Pr\{N(t) = n\}.$$

The Chapman-Kolmogorov equations can be written down as follows:

$$p_0(t+h) = p_0(t)[1-\lambda h] + p_1(h)[1-\lambda h]k\mu h + o(h),$$

and for $n \geq k$,

$$\begin{aligned} p_n(t+h) = {} & p_n(t)[1-\lambda h][1-k\mu h] + p_{n+1}(t)[1-\lambda h][k\mu h] \\ & + p_{n-k}(t)[\lambda h(1-k\mu h)] + o(h); \end{aligned}$$

for $1 \leq n \leq k$, the last term of the preceding will not occur.

Thus, we get, with $p_j(.) = 0$ for $j < 0$,

$$p_0'(t) = -\lambda p_0(t) + k\mu p_1(t) \tag{4.1.1}$$

$$p_n'(t) = -(\lambda + k\mu)\, p_n(t) + k\mu p_{n+1}(t) + \lambda p_{n-k}(t), \quad n \geq 1. \tag{4.1.2}$$

Assume that $\rho = \lambda/\mu < 1$, that the system is in steady state. Let $p_n = \lim_{t\to\infty} p_n(t)$ be the steady-state probability that there are n stages in the total system. Then we have, with $p_j = 0$ for $j < 0$,

$$\lambda p_0 = k\mu p_1 \tag{4.1.3}$$

$$(\lambda + k\mu)\, p_n = k\mu p_{n+1} + \lambda p_{n-k}, \quad n \geq 1. \tag{4.1.4}$$

Let $P(s) = \sum_{n-0}^{\infty} p_n s^n$ be the PGF of $\{p_n\}$. Multiplying (4.1.4) by s^n and adding over all admissible values of n, we get

$$(\lambda + k\mu)\sum_{n=1}^{\infty} p_n s^n = k\mu \sum_{n=1}^{\infty} p_{n+1} s^n + \lambda \sum_{n=k}^{\infty} p_{n-k} s^k \quad \text{or}$$

$$(\lambda + k\mu)[P(s) - p_0] = \frac{k\mu}{s}[P(s) - p_0 - p_1 s] + \lambda s^k P(s),$$

whence we get

$$P(s) = \frac{p_0\left[(\lambda + k\mu) - \frac{k\mu}{s}\right] - k\mu p_1}{(\lambda + k\mu) - \frac{k\mu}{s} - \lambda s^k}.$$

Writing p_1 in terms of p_0 (from (4.1.3)) and simplifying, we get

$$P(s) = \frac{k\mu p_0(1 - s)}{k\mu - (\lambda + k\mu)s + \lambda s^{k+1}}. \tag{4.1.5}$$

We can evaluate p_0 by using the relation $P(1) = 1$. Using L'Hôpital's rule, we get

$$P(1) = \lim_{s \to 1} P(s) = \frac{-k\mu p_0}{-(\lambda + k\mu) + (k + 1)\lambda}$$

so that

$$p_0 = \frac{(\mu - \lambda)}{\mu} = 1 - \rho.$$

Thus, we have

$$P(s) = \frac{k\mu(1 - \rho)(1 - s)}{k\mu - (\lambda + k\mu)s + \lambda s^{k+1}} \tag{4.1.6}$$

$$= \frac{(1 - \rho)(1 - s)}{(1 - s) - \frac{\lambda s}{k\mu}(1 - s^k)}. \tag{4.1.6a}$$

To find p_n we have to expand $P(s)$ as a power series in s. The usual approach is by a partial fraction expansion of $P(s)$ for which one needs to find the zeros of the denominator; from (4.1.6a) we get

$$P(s) = \frac{(1 - \rho)}{1 - \frac{\lambda}{k\mu}\{s + s^2 + \cdots + s^k\}}. \tag{4.1.6b}$$

Let the zeros of the denominator be $s_1, s_2, \ldots, s_k$; the zeros are unique—that is, there is no multiple zero. The denominator of (4.1.6b) can be written as

$$
\begin{aligned}
&-\frac{\lambda}{k\mu}\left\{s^k + \cdots + s - \frac{k\mu}{\lambda}\right\} \\
&\quad = -\frac{\lambda}{k\mu}(s - s_1)\ldots(s - s_k) \quad \left(\text{where } (-1)^k(s_1 \ldots s_k) = -\frac{k\mu}{\lambda}\right) \\
&\quad = -\frac{\lambda}{k\mu}\{(-1)^k(s_1 \ldots s_k)\} \\
&\qquad \times \left\{\left(1 - \frac{s}{s_1}\right)\left(1 - \frac{s}{s_2}\right)\ldots\left(1 - \frac{s}{s_k}\right)\right\} \\
&\quad = \left(1 - \frac{s}{s_1}\right)\ldots\left(1 - \frac{s}{s_k}\right).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
P(s) &= \frac{(1-\rho)}{\left(1 - \frac{s}{s_1}\right)\left(1 - \frac{s}{s_2}\right)\ldots\left(1 - \frac{s}{s_k}\right)} \\
&= (1-\rho)\sum_{i=1}^{k} \frac{a_i}{\left(1 - \frac{s}{s_1}\right)},
\end{aligned} \tag{4.1.7}
$$

where

$$
a_i = \prod_{\substack{m-1 \\ m \neq i}}^{k} \frac{1}{\left(-\frac{s_1}{s_m}\right)}
$$

Thus,

$$
P(s) = (1-\rho)\sum_{i-1}^{k} a_i\left(1 - \frac{s}{s_i}\right)^{-1}
$$

so that p_n, the coefficient of s^n, is given by

$$
p_n = (1-\rho)\left[\sum_{i=1}^{k} a_i (s_i)^{-n}\right], \quad n = 1, 2, \ldots \tag{4.1.8}
$$

Now p_n gives the probability of the number of *stages* in the system. If $p_r^{(c)}$ denotes the probability that the number of *customers* in the system is r, then we have the following relation between $p_n^{(c)}$ and p_n:

$$
p_n^{(c)} = \sum_{m=(n-1)k+1}^{nk} p_m, \quad n = 1, 2, 3, \ldots \tag{4.1.9}
$$

Thus, from p_n, $p_n^{(c)}$ can be obtained by using the preceding relation.

4.1.1.1 Particular case: $M/M/1$

For $k = 1$, we have $E \equiv M$; then the system becomes $M/M/1$. Putting $k = 1$ in (4.1.6a) we have

$$P(s) = \frac{1-\rho}{1-s\rho} = (1-\rho)(1-s\rho)^{-1}$$

so that $p_n = (1-\rho)\rho^n, n = 0, 1, 2, \ldots$. For $k = 1$, we get from (4.1.9)

$$p_r^{(c)} = \sum_{n=r}^{r} p_n = p_r$$

so that the steady-state distribution that there are n in the system is given by

$$p_n^{(c)} = (1-\rho)\rho^n, \quad n \geq 0.$$

Note 1: An alternative approach to find p_n is as follows: From (4.1.6a) we have

$$\begin{aligned}
P(s) &= \frac{(1-\rho)}{1-\frac{\lambda s}{\mu k}\left\{\frac{(1-s^k)}{1-s}\right\}} \\
&= (1-\rho)\left[1-\frac{\lambda s}{\mu k}(1-s^k)(1-s)^{-1}\right]^{-1} \\
&= (1-\rho)\left[\sum_{m=0}^{\infty}\left(\frac{\lambda}{k\mu}\right)^m s^m(1-s^k)^m(1-s)^{-m}\right] \\
&= (1-\rho)\left[\sum_{m=0}^{\infty}\left(\frac{\lambda s}{k\mu}\right)^m\left\{\sum_{i=0}^{m}(-1)^i\binom{m}{i}(s^k)^i\right\}\right. \\
&\quad \left.\times\left\{\sum_{j=0}^{\infty}\binom{m+j-1}{j}s^j\right\}\right] \quad \text{(where } m \geq i, m+j-1 \geq j\text{)} \\
&= (1-\rho)\left[\sum_{m=0}^{\infty}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(-1)^i\left(\frac{\lambda}{k\mu}\right)^m\binom{m}{i}\binom{m+j-1}{j}s^{m+ik+j}\right].
\end{aligned}$$

Put $n = m + ik + j (m \geq i, m \geq 1)$, then

$$\begin{aligned}
p_n &= \text{coefficient of } s^n \text{ in } P(s) \\
&= (1-\rho)\sum_{m,i,j}(-1)^i\left(\frac{\lambda}{k\mu}\right)^m\binom{m+j-1}{j}, \qquad (4.1.10)
\end{aligned}$$

where $m \geq i, m \geq 1, n = m + ik + j$. For example, for $k = 2$,

$$p_7 = (1-\rho)\left(\rho_1^7 + 6\rho_1^6 + 10\rho_1^5 + 4\rho_1^4\right), \quad \text{where} \quad \rho_1 = \left(\frac{\lambda}{2\mu}\right) = \left(\frac{\rho}{2}\right).$$

Note 2: Another model can be considered by specifying the state of the system by the number of customers in the queue (or system) and the number of the stage in which the customer is being served. Let $p_{q,i}(t)$ denote the probability that there are $q(\geq 0)$ in the queue and the customer being served, if any, is at stage $i(\geq 0)$, with $p_{0,0} = p_0$, and let $p_{q,i}$ be the corresponding steady-state probability. Then the differential-difference equations of the system can be easily written down. The equations are not easy to handle, however. The probability that the number in the system is n is given by

$$p_n = \sum_{i-1}^{k} p_{n-1,i}, \quad (n \geq 1).$$

Note 3: We shall consider later the more general system $M/G/1$ from which results of $M/E_k/1$ can be deduced.

4.1.2 The system $E_k/M/1$

Now we consider a single-channel system where the interarrival time has Erlang-k distribution with mean $1/\lambda$ and the service time has exponential distribution with mean $(1/\mu)$. Here the arrival mechanism may be considered as consisting of k independent exponential stages each with mean $(1/k\lambda)$. An arriving customer has to pass through k successive independent stages, and only when he has passed through all the stages is he finally admitted to the system. Further, the next arrival after him, who arrives when the earlier arrival is at some stage of the arriving mechanism, cannot be admitted (to the first stage) and has to remain in queue unless the earlier arrival has completed all the stages and is admitted to the system. The system state at any epoch may be studied in terms of the arrival stage, the stage being marked in the order in which it occurs. Thus, if at any epoch, there are $j(>0)$ customers (already admitted) in the system and the arriving customer is at the ith stage $(1 \leq i \leq k)$ of the arriving mechanism, then the total number of (completed) stages in the total system r is given by

$$r = jk + (i - 1). \tag{4.1.11}$$

Here with the passage of a customer from stage i to stage $i + 1$ of the arrival mechanism, the number of completed stages changes and transition occurs from state r to state $r + 1$, whereas with the completion of service of a customer, the transition occurs from state r to state $r - k$. Assume that $\rho = \lambda/\mu < 1$—that is, the system is in steady state. Let p_n be the steady-state probability that the number of *stages* in the total system is n. If $p_r^{(c)}$ is the probability that the

number of customers in the system is r, then

$$p_n^{(c)} = \sum_{m=nk}^{k(n+1)-1} p_m, \quad n = 1, 2, \ldots. \tag{4.1.12}$$

The Chapman-Kolmogorov equations of the system can be easily written down. The difference equations of the total system in steady state are given by

$$k\lambda p_0 = \mu p_k \tag{4.1.13}$$

$$k\lambda p_n = k\lambda p_{n-1} + \mu p_{n+k}, \quad 1 \le n \le k-1 \tag{4.1.14}$$

$$(k\lambda + \mu) p_n = k\lambda p_{n-1} + \mu p_{n+k}, \quad n \ge k. \tag{4.1.15}$$

Let $P(s) = \sum_{n=0}^{\infty} p_n s^n$ be the PGF of $\{p_n\}$. Multiplying (4.1.14) and (4.1.15) by s^n and adding over all admissible values of n, we get

$$\sum_{n=1}^{\infty} (k\lambda + \mu) p_n s^n - \sum_{n=1}^{k-1} \mu p_n s^n = \sum_{n=1}^{\infty} k\lambda p_{n-1} s^n + \sum_{n=1}^{\infty} \mu p_{n+k} s^n.$$

Expressing the preceding in terms of $P(s)$ and the missing terms, we get

$$(k\lambda + \mu)[P(s) - p_0] - \sum_{n=1}^{k-1} \mu p_n s^n = k\lambda s P(s) + \frac{\mu}{s^k}\left[P(s) - \sum_{n=0}^{k} p_n s^n\right]. \tag{4.1.16}$$

Multiplying by s^k/μ and collecting terms involving $P(s)$ on one side and noting that $\rho = \lambda/\mu$, we get

$$\begin{aligned} &P(s)[k\rho s^{k+1} - (k\rho + 1)s^k + 1] \\ &\quad = -s^k(k\rho + 1) p_0 - s^k \sum_{n=1}^{k-1} p_n s^n + \sum_{n=0}^{k} p_n s^n \\ &\quad = -s^k \sum_{n=0}^{k-1} p_n s^n - k\rho s^k p_0 + \left(p_k s^k + \sum_{n=0}^{k-1} p_n s^n\right) \\ &\quad = (1 - s^k) \sum_{n=0}^{k-1} p_n s^n \quad \text{(using (4.1.13))}. \end{aligned}$$

Thus,

$$P(s) = \frac{(1 - s^k) \sum_{n=0}^{k-1} p_n s^n}{k\rho s^{k+1} - (k\rho + 1)s^k + 1}. \tag{4.1.17}$$

The expression on the RHS involves $\sum_{n=0}^{k-1} p_n s^n$, which must be eliminated. The denominator has $(k + 1)$ zeros, of which 1 is one. By Rouché's theorem,

we find that $(k-1)$ zeros are of modulus less than unity and one zero is of modulus greater than unity. Denote these zeros by $s_1, \ldots, s_{k-1}$ $(|s_i| < 1,\ i = 1, 2, \ldots, k)$ and $s_0 (|s_0| > 1)$. As $P(s)$ is analytic inside $|s| < 1$ and is bounded, the numerator must also have as zeros $s_1, \ldots, s_{k-1}$, which must come from one of the two factors in the numerator. The zeros of the factor $(1 - s^k)$ have all modulus equal to unity, so the factor $\sum_{n=0}^{k-1} p_n s^n$ must have as roots $s_1, \ldots, s_{k-1}$. Thus,

$$\frac{\sum_{n=0}^{k-1} p_n s^n}{k\rho s^{k+1} - (k\rho + 1)s^k + 1} = \frac{A(s - s_1)\ldots(s - s_{k-1})}{(1 - s)(s - s_0)(s - s_1)\ldots(s - s_{k-1})}$$
$$= \frac{A}{(1 - s)(s - s_0)},$$

where A is a constant to be determined. Substituting in (4.1.17), we get

$$P(s) = \frac{(1 - s^k)A}{(1 - s)(s - s_0)}. \tag{4.1.18}$$

Since $P(1) = 1$, we get by using L'Hôpital's rule

$$1 = P(1) = \lim_{s \to 1} P(s)$$
$$= \frac{-kA}{-(1 - s_0)},$$

so that

$$P(s) = \frac{(1 - s^k)(1 - s_0)}{k(1 - s)(s - s_0)} = \frac{(1 - s^k)\left(1 - \frac{1}{s_0}\right)}{k(1 - s)\left(1 - \frac{s}{s_0}\right)}. \tag{4.1.19}$$

To find p_n, we expand the RHS by resolving first into partial fractions.

We have

$$P(s) = \frac{(1 - s^k)}{k}\left[\frac{1}{1 - s} - \frac{\frac{1}{s_0}}{1 - \frac{s}{s_0}}\right]$$
$$= \frac{(1 - s^k)}{k}\left[\sum_{r=0}^{\infty} s^r - \sum_{r=0}^{\infty} \frac{s^r}{(s_0)^{r+1}}\right]. \tag{4.1.20}$$

Thus, for $0 \leq n < k$,

$$p_n = \frac{1}{k}\left\{1 - s_0^{-(n+1)}\right\}$$

and for $n \geq k$,

$$p_n = \frac{1}{k}\{1 - s_0^{-(n+1)} - 1 + (s_0)^{-(n-k+1)}\}$$
$$= \frac{1}{k}\{s_0^{k-n-1}(1 - s_0^{-k})\}.$$

Since s_0 is a root of $k\rho s^{k+1} - (k\rho + 1)s^k + 1 = 0$, we get

$$k\rho(s_0 - 1) = 1 - s_0^{-k}. \tag{4.1.21}$$

Thus, we have

$$p_n = \frac{1}{k}\{1 - s_0^{-(n+1)}\}, \quad 0 \leq n < k$$
$$= \rho s_0^{k-n-1}(s_0 - 1), \quad n \geq k \tag{4.1.22}$$

Using (4.1.12), we get, for $n > 0$,

$$p_n^{(c)} = \sum_{m=nk}^{k(n+1)-1} \rho s_0^{k-m-1}(s_0 - 1)$$
$$= \rho s_0^{k-1}(s_0 - 1)\left[\frac{s_0^{-kn} - s_0^{-k(n+1)}}{1 - s_0^{-1}}\right]$$
$$= \rho s_0^{k-nk}(1 - s_0^{-k}) \tag{4.1.23}$$
$$= (\rho s_0^k)\{(1 - s_0^{-k})(s_0^{-k})^n\}; \tag{4.1.23a}$$

and, for $n = 0$,

$$p_0^{(c)} = \sum_{m=0}^{k-1} p_m = \sum_{m=0}^{k-1} \frac{1}{k}\{1 - s_0^{-(m+1)}\}$$
$$= 1 - \frac{1}{k}\frac{s_0^{-1}(1 - s_0^{-k})}{1 - s_0^{-1}}$$
$$= 1 - \frac{1}{k}\frac{(1 - s_0^{-k})}{(s_0 - 1)}$$
$$= 1 - \rho \quad \text{(using (4.1.21))}. \tag{4.1.24}$$

Thus, we find that the distribution of the number in the system is given by (4.1.23) for $n > 0$ and by (4.1.24) for $n = 0$. The form of the expression (4.1.23a) implies that the distribution of the number of customers in the system is *modified geometric* with a slightly modified first term. We shall take up further questions later on when we study more general $G/M/1$ systems.

Note: The arithmetic-geometric mean inequality to the characteristic equation (denominator of (4.1.17) equated to zero) gives the result that s_0 is real and that

$$1 < s_0 \leq \rho^{-2/(k+1)}.$$

4.2 Bulk Queues

Bailey (1954) introduced the concept of bulk queues; he considered a situation where service can be effected in a batch of up to C customers—that is, all waiting customers up to a fixed capacity C are taken for service in a batch. Gaver (1959) introduced bulk-arrival queues, where arrival could be in bulk or batch. The literature on bulk queues with bulk arrival and/or with bulk service is now quite vast. Chaudhry and Templeton (1983) discuss this subject at great length. See also Medhi (1984, 1994).

We shall first discuss bulk queues that can be modeled as non-birth-and-death processes. The resulting queueing processes will be Markovian. The Chapman-Kolmogorov equations or the balance equations can be written down easily, and the analysis can be done in a straightforward manner similar to that of birth-and-death queueing processes.

4.2.1 Markovian bulk-arrival system: $M^X/M/1$

Let us consider a single-server queueing system in which customers arrive in batches in accordance with a time-homogeneous Poisson process with parameter λ. Assume that the batch size X is a RV with PF

$$Pr(X = k) = a_k, \quad k = 1, 2, 3, \ldots.$$

that is, the probability that a batch of k arrives in an infinitesimal interval $(t, t + h)$ is $\lambda a_k h + o(h)$.

Let $A(s) = \sum_{k=1}^{\infty} a_k s^k$ be the PGF of X and $\bar{a} = A'(1) = E(X)$ be the mean of X. The arrival process is a compound Poisson process with mean arrival rate $\lambda\bar{a}$. Assume that service takes place singly and that the distribution of service time is exponential with mean $1/\mu$. We consider the number N of customers in the system. The underlying process is still Markovian but is a non-birth-death, as transitions, in case of arrival of a batch of size k ($k \geq 1$), will occur to a state differing by k and will not necessarily be to a neighboring one. The system is denoted by $M^X/M/1$. The utilization factor is $\rho = \lambda E(X)/\mu = \lambda\bar{a}/\mu$. The balance equations can be written down as:

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \tag{4.2.1}$$

$$p_n'(t) = -(\lambda + \mu) p_n(t) + \mu p_{n+1}(t) + \sum_{k=1}^{n} \lambda a_k p_{n-k}(t), \quad n \geq k \geq 1. \tag{4.2.2}$$

This queue is considered also in Stadje (1989).

4.2.1.1 Steady-state solution

Assume that the steady-state solution exists. Then we get

$$0 = -\lambda p_0 + \mu p_1 \tag{4.2.3a}$$

$$0 = -(\lambda + \mu) p_n + \mu p_{n+1} + \lambda \sum_{k=1}^{n} a_k p_{n-k}, \quad n \geq k \geq 1. \tag{4.2.3b}$$

Let $P(s) = \sum_{n=0}^{\infty} p_n s^n$ be the PGF of $\{p_n\}$. Multiplying (4.2.3b) by s^n for $n = 1, 2, 3, \ldots$ and adding them to (4.2.3a), we get

$$0 = -\lambda P(s) - \mu[P(s) - p_0] + \left(\frac{\mu}{s}\right)[P(s) - p_0] + \lambda \sum_{n=1}^{\infty} \sum_{k=1}^{n} a_k p_{n-k} s^n. \tag{4.2.4}$$

The last term on the RHS of Eq. (4.2.4) can be written as

$$\lambda \sum_{k=1}^{\infty} a_k s^k \left\{ \sum_{n=k}^{\infty} p_{n-k} s^{n-k} \right\} = \lambda A(s) P(s).$$

Then from (4.2.4) we get

$$P(s) = \frac{\mu(1-s) p_0}{\mu(1-s) - \lambda s[1 - A(s)]}. \tag{4.2.5a}$$

Since $P(1) = 1$, we get, applying L'Hôpital's rule,

$$1 = \lim_{s \to 1} P(s) = \frac{-\mu p_0}{-\mu + \lambda \bar{a}},$$

whence

$$p_0 = 1 - \lambda \frac{\bar{a}}{\mu} = 1 - \rho.$$

The condition $\rho < 1$ is sufficient for existence of steady state. From (4.2.5a) we get

$$P(s) = \frac{\mu(1-s)(1-\rho)}{\mu(1-s) - \lambda s[1 - A(s)]}$$

$$= \frac{(1-\rho)}{1 - \frac{\lambda s}{\mu(1-s)}\{1 - A(s)\}}. \tag{4.2.5b}$$

4.2.1.2 Expected number in the system

The expected number in the system $E(N)$ is given by $E(N) = P'(1)$. To find $P'(1)$, we can proceed in a straightforward manner, or, better still, we can use a method described shortly where $P(s)$ is indeterminate at $s = 1$. To find $P(1)$

we used L'Hôpital's rule, and it was easy. To find $P(1)$ and also $P'(1)$, proceed as follows.

The power series for $P(s)$ is convergent in the full unit interval, and, thus, by Abel's theorem the functions $P(s)$ and $P'(s)$ are continuous at $s = 1$. The indeterminancy in $P(s)$ can be eliminated by expanding the denominator as a power series about the point $s = 1$ and then canceling the factor $(1 - s)$ from the numerator and the denominator of $P(s)$.

Considering the expansion of $f(s) = 1 - C(s)$ about $s = 1$ by Taylor's theorem, we get

$$f(s) = f(1) + (s-1)f'(1) + \frac{(s-1)^2}{2} f''(1) + \text{higher powers of } (s-1).$$

Now

$$f(1) = 0, \qquad f'(1) = -C'(1) = -\bar{a} \quad \text{and}$$
$$f''(1) = -C''(1) = -\sigma^2 + \bar{a}^2 - \bar{a},$$

where $\bar{a}$ is the mean and σ^2 is the variance of the bulk distribution $\{a_k\}$. Thus,

$$\begin{aligned} P(s) &= \frac{\mu(1-s)(1-\rho)}{\mu(1-s) - \lambda s\left[(s-1)f'(1) + \frac{(s-1)^2}{2!} f''(1) + \cdots\right]} \\ &= \frac{\mu(1-\rho)}{\mu + \lambda s f'(1) + \frac{\lambda s(s-1)}{2!} f''(1) + \cdots}. \end{aligned} \tag{4.2.6}$$

It readily follows that

$$P(1) = \frac{\mu(1-\rho)}{\mu - \lambda\bar{a}} = 1$$

$$\begin{aligned} P'(1) &= \frac{\mu(1-\rho)}{(\mu - \lambda\bar{a})^2}\left[-\lambda f'(1) - \frac{\lambda}{2} f''(1)\right] \\ &= \frac{1}{\mu - \lambda\bar{a}}\left[\lambda C'(1) + \frac{\lambda}{2} C''(1)\right] \\ &= \frac{1}{\mu - \lambda\bar{a}}\left[\lambda\bar{a} + \frac{\lambda}{2}(\sigma^2 + \bar{a}^2 - \bar{a})\right]. \end{aligned}$$

Thus,

$$E(N) = \frac{\lambda(\sigma^2 + \bar{a}^2 + \bar{a})}{2(\mu - \lambda\bar{a})} = \frac{\rho}{1-\rho}\left\{\frac{E(X^2) + E(X)}{2E(X)}\right\}, \tag{4.2.7}$$

which gives the expected number in the system. In the same manner $P''(1)$ can be found, and the variance of the number in the system can be obtained.

Writing $E(X^k) = a^{(k)}, k \geq 2, a^{(1)} = \bar{a}$, we get

$$E\{N\} = \frac{\rho}{1-\rho}\left\{\frac{a^{(2)} + \bar{a}}{2\bar{a}}\right\}$$

and

$$\text{var}\{N\} = \frac{\rho}{1-\rho}\left\{\frac{2a^{(3)} + 3a^{(2)} + \bar{a}}{6\bar{a}}\right\} + [E\{N\}]^2.$$

4.2.1.3 Particular cases

(1) By considering that $a_k = 0, k \neq 1, a_1 = 1$, we get the corresponding result for the $M/M/1$ queue.

(2) Let X have geometric (decapitated) distribution

$$a_k = P(X = k) = c(1-c)^{k-1}, \quad 0 < c < 1, \quad k = 1, 2, \ldots,$$

with PGF $A(s) = cs/[1-(1-c)s]$ and $\bar{a} = E(X) = A'(1) = 1/c$. Then

$$\begin{aligned} P(s) &= \frac{\mu(1-\rho)(1-s)}{\mu(1-s) - \{s\lambda(1-s)\}/\{1-(1-c)s\}} \\ &= \frac{(1-\rho)\{1-(1-c)s\}}{1-\{(1-c)+c\rho\}s}, \quad \text{since } \rho = \lambda/c\mu \\ &= (1-\rho)\left[\frac{1}{1-\{(1-c)+c\rho\}s}\,\frac{(1-c)s}{1-\{(1-c)+c\rho\}s}\right] \\ &= (1-\rho)\left[\sum_{n=0}^{\infty}\{(1-c)+c\rho\}^n s^n \right. \\ &\qquad \left. -(1-c)\sum_{n-0}^{\infty}\{(1-c)+c\rho\}^n s^{n+1}\right] \end{aligned}$$

so that

$$\begin{aligned} p_n &= (1-\rho)[\{(1-c)+c\rho\}^n - (1-c)\{(1-c)+c\rho\}^{n-1}] \\ &= c\rho(1-\rho)[(1-c)+c\rho]^{n-1}, \quad n \geq 1 \\ &= \rho\{c(1-\rho)\}\{1-c(1-\rho)\}^{n-1}. \end{aligned}$$

This, together with $p_0 = 1-\rho$, gives $\{p_n\}$. The distribution is geometric.

(3) Let the batch be of size 1 or 2 with equal probability

$$a_1 = a_2 = \frac{1}{2}, \qquad a_k = 0, \quad k > 2;$$

$$A(s) = \frac{1}{2}s(1+s), \qquad \rho = \frac{2}{3}\frac{\lambda}{\mu}.$$

Then

$$P(s) = \frac{\mu(1-\rho)(1-s)}{\mu(1-s) - \lambda s\left\{1 - \frac{1}{2}s(1+s)\right\}} = \frac{3(1-\rho)}{6 - 4\rho s - 2\rho s^2}$$
$$= \frac{3(1-\rho)}{(s_2 - s_1)}\left[\frac{1}{s - s_1} - \frac{1}{s - s_2}\right],$$

where s_1 and s_2 are the roots of the equation $(3/\rho) - 2s - s^2 = 0$. The RHS can be expanded in powers of s, and from there an expression for p_n can be obtained.

A more general case where X has a Bernoulli distribution with $a_1 = p, a_2 = 1 - p, 0 < p < 1$, can be similarly treated. Jensen *et al.* (1977) consider the case where X has a multinomial distribution.

(4) For fixed batch size $M^r/M/1$, suppose each batch consists of exactly r arrivals. Then $a_r = 1, a_k = 0, k \neq r$ and $A(s) = s^r$. The PGF of the number of customers in the system $M^r/M/1$ is [from (4.2.5a)],

$$P(s) = \frac{\mu(1-s)(1-\rho)}{\mu(1-s) - \lambda s(1-s^r)}, \tag{4.2.8}$$

where $\rho = r\lambda/\mu$.

4.2.2 Equivalence of $M^r/M/1$ and $M/E_r/1$ systems

In the system $M/E_r/1$, arrivals occur singly in accordance with a Poisson process, while the service consists of r stages, each exponential with mean $(1/r\mu)$ (parameter $r\mu$), and the completion of the service requires total service in r stages with mean $r(1/r\mu) = 1/\mu$. The system can be considered as one in which "each customer" arrival amounts to arrival of "r customers," such that each of these would require a single stage of exponential service with mean $(1/r\mu)$. This latter system is $M^r/M/1$ with mean service time $(1/r\mu)$. Thus, the distribution of the number of customers in the system $M^r/M/1$ (with μ replaced by μr) is the same as the distribution of the number of stages in the system $M/E_r/1$. The equivalence of the two systems is thus established. Replacing μ by $r\mu$ in (4.2.8), we get the PGF of the number of stages in $M/E_r/1$, which is given by

$$P(s) = \frac{r\mu(1-s)(1-\rho)}{r\mu(1-s) - \lambda s(1-s^r)}, \quad (\rho = \lambda/\mu). \tag{4.2.9}$$

4.2.3 Waiting-time distribution in an $M^X/M/1$ queue

The waiting time of a test unit consists of two components:

(1) the time required to complete service of all the units in the system found by an arriving group (call this delay D_1), and

(2) the time to serve all units of the group (in which the test unit arrived) who are served prior to the start of service of the test unit (call this delay D_2).

The total waiting (queueing) time D of a test unit is the sum of these two independent RVs—that is, $D = D_1 + D_2$. This has been discussed in detail for the more general $M^X/G/1$ model in Section 6.7.2. One can follow the same arguments to find the LST of the distributions of D_1, D_2, D *or* simply get the LSTs by putting the LST of exponential service time with rate μ, i.e. $B^*(s) = \frac{\mu}{s+\mu}$ in the expressions (6.7.5), (6.7.8), and (6.7.9) respectively (as a particular case). The moments of D_1, D_2, D can be obtained for this particular case from there given in Section 6.7.2.2. Writing $E(X^k) = a^{(k)}, k \geq 2, E(X) = \bar{a}$, we have, for $M^X/M/1$,

$$E(D_1) = \frac{\lambda}{2\mu^2(1-\rho)} \left[a^{(2)} + \bar{a}\right] \tag{4.2.10}$$

$$\text{and} \quad E(D_2) = \frac{1}{2\mu}\left[\frac{a^2}{\bar{a}} - 1\right] = \frac{1}{2\mu\bar{a}}\left[a^{(2)} + \bar{a}\right] - \frac{1}{\mu} \tag{4.2.11}$$

Thus, the expected waiting time in the system (response time—that is, queueing plus service time) of a test unit is given by

$$\begin{aligned} E(W) &= E(D_1) + E(D_2) + \frac{1}{\mu} \\ &= \frac{1}{2\mu(1-\rho)}\left[\frac{a^{(2)} + \bar{a}}{\bar{a}}\right]. \end{aligned} \tag{4.2.12}$$

Using Little's Law, we get $E\{N\} = (\lambda\bar{a})E(W)$; this agrees with (4.2.7). (See treatment of $M^X/G/1$ in Section 6.5. for more discussion).

4.2.4 Transient-state behavior

Let us consider now the transient state of the system. Assume that $p_0(0) = 1$—that is, at time 0 there is no customer in the system. Let

$$p_n^*(\alpha) = \text{LT of } p_n(t) = \int_0^\infty e^{-\alpha t} p_n(t)dt \quad \text{and} \tag{4.2.13}$$

$$P^*(s,\alpha) = \sum_{n=0}^{\infty} p_n^*(\alpha)s^n \tag{4.2.14}$$

be the GF of $\{p_n^*(\alpha)\}$. Taking the LT of (4.2.1) and (4.2.2) we get

$$\alpha p_0^*(\alpha) - 1 = -\lambda p_0^*(\alpha) + \mu p_1^*(\alpha) \tag{4.2.15}$$

$$\alpha p_n^*(\alpha) = -(\lambda+\mu)p_n^*(\alpha) + \mu p_{n+1}^*(\alpha) + \sum_{k=1}^{\infty} \lambda a_k p_{n-k}^*(\alpha), \quad n \geq k \geq 1. \tag{4.2.16}$$

Proceeding in the same manner, we get on simplification

$$P^*(s,\ \alpha) = \frac{s + \mu(s-1)p_0^*(\alpha)}{s(\alpha + \lambda + \mu) - \mu - \lambda s A(s)}. \tag{4.2.17}$$

It can be easily verified by taking limits that the corresponding steady-state result follows. We have

$$\begin{aligned} P(s) &= \sum_{n=0}^{\infty} p_n s^n = \sum_{n=0}^{\infty} \left\{ \lim_{t\to\infty} p_n(t) \right\} s^n \\ &= \sum_{n=0}^{\infty} \lim_{\alpha\to 0} \alpha p_n^*(\alpha) s^n \\ &= \lim_{\alpha\to 0} \alpha P^*(s,\alpha), \end{aligned} \tag{4.2.18}$$

as can be verified; $p_0^*(\alpha)$ can be evaluated by the method discussed by Luchak (1958).

Note: The initial condition $p_i(0) = 1$ will yield the same form as (4.2.17) with s^{i+1} in place of the first term s in the numerator of (4.2.17).

4.2.4.1 Busy-period distribution

Let $N^*(t)$ be the number in the system during a busy period starting with one customer at $t = 0$. Let

$$q_n(t) = P\{N^*(t) = n \mid N^*(0) = 1\}. \tag{4.2.19}$$

Then $q_1(0) = 1, q_n(0) = 0, n \geq 2$. The equations governing the zero-avoiding state probabilities will be

$$\begin{aligned} q_n'(t) &= -(\lambda + \mu)q_n(t) + \mu q_{n+1}(t) + \lambda \sum_{k=2}^{n} a_k q_{n-k}(t), \quad n \geq k \geq 2 \\ q_1'(t) &= -(\lambda + \mu)q_1(t) + \mu q_2(t). \end{aligned} \tag{4.2.20}$$

The GF of the LT of $q_n(t)$ can be obtained, and from there the LT $q_1^*(s)$ of the busy-period density can be found.

Mean Busy Period

Let

$$B = \text{length of busy period}$$
$$I = \text{length of idle period}$$

We have $E(I) = 1/\lambda$,

$$\frac{E(B)}{E(I)+E(B)} = \rho = \frac{\lambda E(X)}{\mu},$$

and $$\frac{E(I)}{E(I)+E(B)} = 1-\rho.$$

From the above, we get

$$E(B) = \frac{\rho}{1-\rho}.E(I) = \frac{E(X)}{\mu(1-\rho)}.$$

4.2.5 The system $M^X/M/\infty$

The transient state was considered by Reynolds (1968). Let $N(t)$ be the number in the system (or the number of busy servers) at time t. We suppose that $\{N(0) = i\}$—that is, there are i in the system at time 0. Let

$$p_{i,j}(t) = P\{N(t) = j \mid N(0) = i\} \quad \text{and} \tag{4.2.21}$$

$$Q(s,t) = \sum_{j=0}^{\infty} p_{i,j}(t)s^j \tag{4.2.22}$$

be the PGF of $\{p_{i,j}(t)\}$. We have $Q(s,0) = s^i$. Now $p_{i,j}(t)$ satisfies the following differential-difference equations:

$$p'_{i,j}(t) = \lambda \sum_{r=1}^{j} a_r\, p_{i,j-r}(t) - (\lambda + j\mu)\, p_{i,j}(t) + (j+1)\mu p_{i,j+1}(t), \quad j = 0,1,2,\ldots \tag{4.2.23}$$

Multiplying (4.2.23) by s^j and summing over all j, we get

$$\frac{\partial Q}{\partial t} = \mu(1-s)\frac{\partial Q}{\partial s} - \lambda[\{1 - A(s)\}Q]. \tag{4.2.24}$$

This is an equation of the form

$$p\frac{\partial Q}{\partial t} + q\frac{\partial Q}{\partial s} = r,$$

where p, q, and r are functions of t, s, and Q. Equation (4.2.24) can be solved by the method of Lagrange, which uses the subsidiary equation

$$\frac{dt}{1} = \frac{ds}{-\mu(1-s)} = \frac{dQ}{\lambda\{1 - A(s)\}Q}. \tag{4.2.25}$$

From the first two, we get, by integration,

$$(s-1)e^{-\mu t} = C_1, \tag{4.2.26}$$

where C_1 is a constant; and from the last two, by integration, we get

$$\int \frac{dQ}{Q} = \frac{\lambda}{\mu}\int \left(\frac{1-A(s)}{1-s}\right) ds \tag{4.2.27}$$

$$= \frac{\lambda}{\mu}\int A_1(s)ds, \tag{4.2.27a}$$

where

$$A_1(s) = \frac{1-A(s)}{1-s} = \sum_{r=0}^{\infty} Pr(X > r)s^r \tag{4.2.27b}$$

is the generating function of $\{Pr(X > r)\}$. Integration of (4.2.27) gives

$$Q\exp\left[-\frac{\lambda}{\mu}\int_0^s A_1(s)ds\right] = C_2, \tag{4.2.28}$$

where C_2 is a constant. The constants C_1 and C_2 can be functionally related by writing $C_2 \equiv F(C_1)$ where F is an arbitrary function. Thus, from (4.2.26) and (4.2.28) we get

$$Q(s,t)\exp\left[-\frac{\lambda}{\mu}\int_0^s A_1(s)ds\right] = F[(s-1)e^{-\mu t}]. \tag{4.2.29}$$

Thus, the general solution of (4.2.24) is given by

$$Q(s,t) = \left\{\exp\left[\frac{\lambda}{\mu}\int_0^s A_1(y)dy\right]\right\}\{F[(s-1)e^{-\mu t}]\}. \tag{4.2.29a}$$

To determine F, we put $t = 0$; LHS becomes $Q(s,0) = s^i$, so that

$$s^i = \left[\exp\left\{\frac{\lambda}{\mu}\int_0^s A_1(y)dy\right\}\right] F[(s-1),$$

whence

$$F(s) = (s+1)^i \exp\left\{-\frac{\lambda}{\mu}\int_0^{s+1} A_1(y)dy\right\}. \tag{4.2.30}$$

Put

$$e^{-\mu t} = p, \qquad 1 - e^{-\mu t} = q; \tag{4.2.31}$$

then

$$(s-1)e^{-\mu t}+1 = sp+q \quad \text{and}$$

$$F\{(s-1)e^{-\mu t}\} = F\{(sp+q-1)\}$$

$$= (sp+q)^i \exp\left\{-\frac{\lambda}{\mu}\int_0^{sp+q} A_1(y)dy\right\}. \quad (4.2.32)$$

Thus, from (4.2.29a) and (4.2.32) we get

$$Q(s,t) = (sp+q)^i \exp\left\{\frac{\lambda}{\mu}\left[\int_0^s A_1(y)dy - \int_0^{sp+q} A_1(y)dy\right]\right\} \quad (4.2.33)$$

$$= (sp+q)^i \exp\left\{-\frac{\lambda}{\mu}\int_s^{sp+q} A_1(y)dy\right\}. \quad (4.2.33a)$$

Thus, we obtain the PGF $Q(s,t)$ of $\{p_{i,j}(t)\}$. When $i=0$, that is,

$$P\{N(0)=0\} = 1,$$

then

$$Q(s,t) = \exp\left\{-\frac{\lambda}{\mu}\int_s^{sp+q} A_1(y)dy\right\}. \quad (4.2.34)$$

4.2.5.1 Steady-state result

As $t \to \infty$, $p = e^{-\mu t} \to 0$, and $q = 1 - e^{-\mu t} \to 1$, $Q(s,t) \to Q(s)$ so that from (4.2.33a) we get

$$Q(s) = \sum_{j=0}^{\infty} p_{i,j} = \exp\left\{-\frac{\lambda}{\mu}\int_s^1 A_1(y)dy\right\}, \quad (4.2.35)$$

which is independent of the initial state i, so that $p_{i,j} = p_j$, $j = 0, 1, 2, \ldots$. Writing

$$\int_s^1 A_1(y)dy = \int_0^1 A_1(y)dy - \int_0^s A_1(y)dy$$

$$= k(1-G(s)), \quad (4.2.36)$$

where

$$k = \int_0^1 A_1(y)dy, \qquad G(s) = \frac{1}{k}\int_0^s A_1(y)dy,$$

we get

$$G(1) = \frac{1}{k}\int_0^1 A_1(y)dy = \frac{1}{k}k = 1,$$

so that $G(s)$ is a PGF. It is the PGF of the distribution

$$\{f_r\} \quad \text{where} \quad f_r = \frac{P(X > r-1)}{kr}, \quad r = 1, 2, 3, \ldots$$

Thus, from (4.2.35) and (4.2.36) we get that

$$Q(s) = \exp\left\{-\frac{\lambda k}{\mu}[1 - G(s)]\right\} \tag{4.2.37}$$

so that the distribution of $\{p_j\}$ is compound Poisson.

4.2.5.2 Particular case $M/M/\infty$: transient state

Let arrivals occur singly; the $P(X = 1) = 1$, $A(s) = s$, and $A_1(s) = 1$. Then from (4.2.33) we get

$$\begin{aligned} Q(s,t) &= (sp+q)^i \exp\left\{-\frac{\lambda}{\mu}(sp+q-2)\right\} \\ &= (sp+q)^i \exp\left\{-\frac{\lambda q}{\mu}(1-s)\right\} \\ &= (sp+q)^i \exp\left\{\frac{\lambda}{\mu}(1-e^{-\mu t})(s-1)\right\}. \end{aligned} \tag{4.2.38}$$

With $i = 0$, we get

$$Q(s,t) = \exp\left\{\frac{\lambda}{\mu}(1-e^{-\mu t})(s-1)\right\} \tag{4.2.39}$$

so that the distribution of $\{p_{0,j}(t)\}$ is Poisson with mean $(\lambda/\mu)(1 - e^{-\mu t})$. Thus,

$$p_j(t) = \left\{\exp\left[-\frac{\lambda}{\mu}(1-e^{-\mu t})\right]\right\} \frac{\left[\frac{\lambda}{\mu}(1-e^{-\mu t})\right]^j}{j!}. \tag{4.2.40}$$

This result is obtained later as a particular case of $M/G/\infty$. Evidently, in steady state, the distribution of $\{p_j\}$ is Poisson with mean λ/μ, and this happens irrespective of the magnitude of λ/μ.

4.2.5.3 Regression of $N(t)$ on $N(0)$

Reynolds (1968) gives another interesting result for the $M^X/M/\infty$ system. Suppose $i > 0$. From (4.2.33) we get the expected number in the system, given $N(0) = i$. We have

$$E\{N(t) \mid N(0) = i\} = \frac{\partial}{\partial s} Q(s, t)|_{s=1}.$$

Noting that

$$\frac{d}{ds}\int_0^s A_1(y)dy = A_1(s),$$

$$\frac{d}{ds}\int_0^{sp+q} A_1(y)dy = \left[\frac{d}{ds}(sp+q)\right] A_1(sp+q)$$

$$= pA_1(sp+q), \quad \text{and}$$

$$\frac{d}{ds}\left\{\exp\frac{\lambda}{\mu}\left[\int_0^s A_1(y)dy - \int_0^{sp+q} A_1(y)dy\right]\right\}\Bigg|_{s=1}$$

$$= \frac{\lambda}{\mu}[A_1(1) - pA_1(1)]\exp\left\{\frac{\lambda}{\mu}\left[\int_0^1 A_1(y)dy - \int_0^1 A_1(y)dy\right]\right\}$$

$$= \frac{\lambda}{\mu} q A_1(1)$$

$$= \frac{\lambda}{\mu} q E(X),$$

we get

$$E\{N(t) \mid N(0) = i\} = ip + \frac{\lambda}{\mu} q A_1(1)$$

$$= \frac{\lambda}{\mu}(1 - e^{\mu t})E(X) + N(0)e^{-\mu t}.$$

This shows that the regression of $N(t)$ on $N(0)$ is linear, being of the form $E\{Y \mid x\} = a + bx$.

4.3 Queueing Models with Bulk (Batch) Service

We have so far considered models with individual service. Now we shall consider systems where services are offered in batches instead of personalized service of one at a time. Bailey (1954) was the first to consider bulk service. The literature on bulk service has grown over the years. Such models find applications in

several situations, such as transportation. Mass transit vehicles and carriers are natural batch servers.

A number of policies of bulk service are considered in literature. These are discussed below.

(1) Bailey (1954) considered that the server serves in batches of size not more than, say, b, the (maximum) capacity of the server. If the server, on completion of a batch service, finds not more than b waiting, then he takes all of them in a batch for service. If he finds more than b waiting, then he takes for service a batch of a size b (in order of arrival or in any other order), while others, in excess of b units, wait and join the queue. An example of this type of server is an elevator.

(2) A service batch may be of a fixed size—say, k. The server waits until there are k in the queue and starts service as soon as the queue reaches this size. If, on completion of a batch service, he finds more than k waiting, the server takes a batch of size k (in order of arrival or in any other order), while others, in excess of k units, wait in the queue.

(3) A server may take in a batch a minimum number of units—say, a less than or equal to his capacity—say, b. The server adopts the following policy. If, on completion of a batch service, he finds q units waiting and if

(i) $0 \leq q < a$, then he waits till the queue size grows to a,

(ii) $a \leq q \leq b$, then he takes a batch of size q for service, and

(iii) $q > b$, then he takes a batch of size b for service (in order of arrival or in random order), while those in excess of b units wait in the queue.

We shall call this rule the *general bulk service rule*, as rules under (1) and (2) above can be covered as particular cases of this rule. Neuts (1967) considered this rule; this has been further investigated, for example, by Medhi, Borthakur, Sim, Templeton, Chaudhry, Powell, Humblet, Bertsimas and others. Kosten earlier considered such a situation with infinite server capacity.

(4) The size of a batch may be a random variable; it may depend on the unfilled capacity of the server. This rule has been considered, for example, by Cohen, Prabhu, Bhat, Teghem, and others. Newell considers a model such that batch service may be extended to accommodate additional units during the course of the service to the extent of the unfilled capacity (of the server or service channel), if there be any. This may be called accessible batch service. For references, see Medhi (1984).

We shall examine here the general bulk service, which, though considered in literature, has not been treated in standard textbooks.

4.3.1 The system $M/M(a, b)/1$

We assume that the input is Poisson (with single arrivals at each epoch of Poisson occurrence) with rate λ. The service is in batches under the general bulk service rule as discussed above. Service commences only when the queue size reaches

or exceeds a, the capacity being $b(\geq a \geq 1)$. The service time distribution of a batch is assumed to be exponential with parameter μ. For simplicity, the service time is taken to be independent of the batch size. We denote the system by the notation $M/M(a, b)/1$. Denote the states of the system by (i, n), where i is an indicator variable. $i = 1$ implies that the server is busy in serving a batch of size s $(a \leq s \leq b)$, and $i = 0$ implies that the server is idle, with n being the number of units in the queue. We consider the states of the system. Denote $p_{i,n}(t) = Pr\{\text{at time } t, \text{ the system is at state } (i, n)\}$. $p_{i,n}(t)$ is non-zero only for $i = 1, n \geq 0$, and $i = 0, 0 \leq n \leq a - 1$.

It is clear that the system is Markovian, arising out of non-birth-death processes. The Chapman-Kolmogorov equations lead to

$$p'_{1,n}(t) = -(\lambda + \mu)\, p_{1,n}(t) + \lambda p_{1,n-1}(t) + \mu p_{1,n+b}(t), \quad n = 1, 2, \ldots \tag{4.3.1}$$

$$p'_{1,0}(t) = -(\lambda + \mu)\, p_{1,0}(t) + \lambda p_{0,a-1}(t) + \mu \sum_{r=a}^{b} p_{1,r}(t) \tag{4.3.2}$$

$$p'_{0,0}(t) = -\lambda p_{0,0}(t) + \mu p_{1,0}(t) \tag{4.3.3}$$

$$p'_{0,q}(t) = -\lambda p_{0,q}(t) + \lambda p_{0,q-1}(t) + \mu p_{1,q}(t), \quad q = 1, 2, \ldots, a - 1. \tag{4.3.4}$$

It is to be noted that (4.3.4) will not occur when $a = 1$.

4.3.1.1 Steady-state solution

Assume that steady state exists. Let

$$p_{0,q} = \lim_{t \to \infty} p_{0,q}(t)$$

$$p_{1,n} = \lim_{t \to \infty} p_{1,n}(t).$$

The steady-state equations of the system become

$$0 = -(\lambda + \mu)\, p_{1,n} + \lambda p_{1,n-1} + \mu p_{1,n+b}, \quad n = 1, 2, \ldots, \tag{4.3.5}$$

$$0 = -(\lambda + \mu)\, p_{1,0} + \lambda p_{0,a-1} + \mu \sum_{r=a}^{b} p_{1,r} \tag{4.3.6}$$

$$0 = -\lambda p_{0,0} + \mu p_{1,0} \tag{4.3.7}$$

$$0 = -\lambda p_{0,q} + \lambda p_{0,q-1} + \mu p_{1,q}, \quad q = 1, 2, \ldots, a - 1. \tag{4.3.8}$$

Equation (4.3.8) will not occur when $a = 1$.

Solution of the preceding difference equations would give the probabilities $p_{0,q}$, $p_{1,n}$. Denoting the displacement operator by E (i.e., $E\{p_{1,r}\} = p_{1,r+1}$),

Eq. (4.3.5) can be written as

$$-(\lambda+\mu)Ep_{1,n-1}+\lambda p_{1,n-1}+\mu E^{b+1}\{p_{1,n-1}\}=0$$

$$\text{or}\quad h(E)\{p_{1,n}\}=0,\quad n=0,1,2,\ldots,$$

with characteristic equation

$$h(z)\equiv \mu z^{b+1}-(\lambda+\mu)z+\lambda=0. \tag{4.3.9}$$

Suppose that $f(z)=-(\lambda+\mu)z$ and $g(z)=\mu z^{b+1}+\lambda$. Consider the circle $|z|=1-\delta$, where δ is arbitrarily small. Writing $z=(1-\delta)e^{i\theta}$, it can be shown that on the contour of the circle,

$$|g(z)|<|f(z)|.$$

Hence from Rouché's theorem it follows that $f(z)$ and $f(z)+g(z)$ will have the same number of zeros inside $|z|=1-\delta$. Since $f(z)$ has only one zero inside this circle, $f(z)+g(z)\equiv h(z)$ will also have only one zero inside $|z|=1-\delta$. This root of $h(z)=0$ is real and unique if and only if $\rho=\lambda/b\mu<1$. Denote this real root by r $(0<r<1)$ and other b roots by $r_1,\ldots,r_b$, $|r_i|\geq 1$.

Then r satisfies the equation

$$b\rho=\frac{\lambda}{\mu}=\frac{r(1-r^b)}{1-r}=r+r^2+\cdots+r^b. \tag{4.3.10}$$

We have from (4.3.10), when $0<\rho<1$

$$\rho\leq r\leq \rho^{2/(b+1)},$$

which is useful in solving Eq. (4.3.10) for r. The solution of (4.3.5) can now be written as

$$p_{1,n}=Ar^n+\sum_{i=1}^{b}A_i r_i^n,\quad n=0,1,2,\ldots$$

where the A's are constants. Since $\sum_{n=0}^{\infty}p_{1,n}<1$ we must have $A_i=0$ for all i, so that, for $n=0,1,2,\ldots$,

$$\begin{aligned}p_{1,n}&=Ar^n\\&=p_{1,0}r^n\\&=\left(\frac{\lambda}{\mu}\right)p_{0,0}r^n\quad\text{from (4.3.7)}\\&=\left(\frac{1-r^b}{1-r}\right)p_{0,0}r^{n+1}.\end{aligned} \tag{4.3.11}$$

Then from (4.3.6), using (4.3.11) and on simplifying, we get

$$p_{0,a-1} = \frac{1-r^a}{1-r} p_{0,0}. \tag{4.3.12}$$

Finally, putting $q = a-1, a-2, \ldots 1$ and using (4.3.11) we get, recursively, for $q = a-2, a-3, \ldots, 1$,

$$p_{0,q} = \frac{1-r^{q+1}}{1-r} p_{0,0} \tag{4.3.13}$$

and because (4.3.12) holds, (4.3.13) holds for $q = 1, 2, \ldots, a-1$. Using

$$\sum_{q=0}^{a-1} p_{0,q} + \sum_{n=0}^{\infty} p_{1,n} = 1,$$

we get

$$p_{0,0} = \left[\frac{a}{1-r} + \frac{r^{a+1} - r^{b+1}}{(1-r)^2}\right]^{-1}. \tag{4.3.14}$$

Thus, (4.3.11), (4.3.13), and (4.3.14) give the steady-state probabilities. The expected number in the queue $E(Q)$

$$\begin{aligned} E(Q) &= \sum_{q=0}^{a-1} q p_{0,q} + \sum_{n=0}^{\infty} n p_{1,n} \\ &= \frac{p_{0,0}}{1-r} \sum_{q=0}^{a-1} \{q - q r^{q+1}\} + \frac{p_{0,0}(1-r^b)}{1-r} \sum_{n=0}^{\infty} n r^{n+1} \\ &= \frac{p_{0,0}}{1-r} \left\{\frac{a(a-1)}{2} + \frac{r^2[a r^{a-1}(1-r) - (1-r^a)]}{(1-r)^2}\right\} \\ &\quad + \frac{p_{0,0}(1-r^b)}{1-r} \frac{r^2}{(1-r)^2}. \end{aligned} \tag{4.3.15}$$

4.3.1.2 Particular cases

(1) Fixed batch size $M/M(k,k)/1$.
Here $a = b = k$, $\rho = \lambda/k\mu$, and r is the real root lying in (0,1) of

$$\frac{\lambda}{\mu} = r + r^2 + \cdots + r^k.$$

Let p_n be the probability that the number in the system N is n. Then

$$p_0 = p_{0,0} = \frac{1-r}{k}$$

$$p_q = p_{0,q} = \frac{1-r^{q+1}}{1-r}p_0 = \frac{1-r^{q+1}}{k}, \quad q = 1, 2, \ldots, k-1,$$

$$p_{m+k} = p_{1,m} = \left(\frac{\lambda}{\mu}\right)p_0 r^m = \rho(1-r)r^m = p_k r^m, \quad m = 0, 1, \ldots.$$

Or, in compact form,

$$p_n = \begin{cases} \frac{1-r^{n+1}}{k}, & n = 0, 1, 2, \ldots, k-1 \\ \rho(1-r)r^{n-k} = p_k r^{n-k}, & n = k, k+1, \ldots \end{cases} \tag{4.3.16}$$

(2) $M/M/1$: The distribution of the number in the system N is obtained by putting $a = b = 1$, then $r = \rho$ and

$$p_0 = P(N = 0) = p_{0,0} = 1 - \rho$$
$$p_n = P(N = n) = p_{1,n-1} = (1-\rho)\rho^n, \quad n = 1, 2, \ldots$$

(3) $M/M(1, b)/1$: Usual bulk service rule.

(4) $M/M(a, \infty)/1$: Here the server has infinite capacity. As $b \to \infty$, $r^b \to 0$ and $\lambda/\mu \to r/(1-r)$, so that $r = \lambda/(\lambda+\mu)$ and

$$p_{0,0} = \left[\frac{a}{1-r} + \frac{r^{a+1}}{(1-r)^2}\right]^{-1}.$$

Note: Neuts and others after him used analytic methods to study systems under this rule. Later, Neuts considered an altogether different approach—an algorithmic approach for queues that have a modified matrix-geometric structure. This method uses a matrix method as an alternative to closed-form analytic methods. For a description of this method, refer to Neuts (1979, 1981).

4.3.2 Distribution of the waiting-time for the system $M/M(a, b)/1$

Medhi (1975) obtained the waiting-time distribution. Assume that the system is in steady state. Let the random variable W_q denote the waiting time in the queue for an arriving unit. Denote

$$\omega(t) = \text{the PDF of } W_q$$
$$f(\alpha, k; t) = \text{the PDF of gamma distribution with parameters } \alpha, k$$
$$= \alpha^k t^{k-1}\frac{\exp(-\alpha t)}{\Gamma(k)}, \quad t > 0, \quad k = 1, 2, \ldots.$$

$$\Gamma_x(\alpha, k) = \int_0^x f(\alpha, k : t)dt$$
$$= 1 - \sum_{r=0}^{k-1} e^{-\alpha x}\frac{(\alpha x)^r}{r!} \tag{4.3.17}$$
$$e(m, z) = \sum_{k=0}^{m-1} \frac{z^k}{k!} \quad \text{and}$$
$$E(m, z) = e^{-z}e(m, z) = \sum_{k=0}^{m-1} \frac{e^{-z}z^k}{k!};$$

$E(m, z) \equiv Pr(X \leq m - 1)$ is the DF of Poisson RV X with mean $z(>0)$. We have then

$$\sum_{q=0}^{a-2} r^q f(\lambda, a - q - 1; t) = \lambda r^{a-2} \exp(-\lambda t) e\left(a - 1, \frac{\lambda t}{r}\right). \tag{4.3.18}$$

The states (sets of) in which a test unit, on arrival, may find the system are

$$\left.\begin{array}{ll} \text{(i)} & (0, a - 1) \\ \text{(ii)} & (0, q), \quad 0 \leq q \leq a - 2 \\ \text{(iii)} & (1, n), \quad a - 1 \leq m \leq b - 1 \\ \text{(iv)} & (1, n), \quad 0 \leq m \leq a - 2 \end{array}\right\} \quad \begin{array}{l} n = kb + m \\ k = 0, 1, 2 \end{array} \tag{4.3.19}$$

If he (or she) finds the system in state (i), then the unit does not wait. Thus, $Pr\{W_q = 0\} = p_{0,a-1}$; the probability of blocking (or delay) equals $Pr\{W_q > 0\} = 1 - p_{0,a-1}$.

If he finds the system in a state of (ii), the unit has to await the arrival of $a - 1 - q$ units after him. The waiting time—that is, the time required for $a - 1 - q$ arrivals—has a gamma distribution with parameters $\lambda, a - 1 - q$.

If he finds the system in a state of (iii), then he has to wait for the completion of services of $k + 1$ batches, including that of the one under service. The time required for completion of services of $k + 1$ batches has a gamma distribution with parameters $\mu, k + 1$.

Consider, finally, that the test unit finds the system in a state of (iv); then he has to wait until either the services of $k + 1$ groups are completed or $a - 1 - m$ units arrive, whichever occurs later. The waiting time in this case is a RV Z, which is the maximum of two gamma variates—that is,

$$Z = \max\{\text{gamma variate with parameters } \lambda, a - 1 - m;$$
$$\text{gamma variate with parameters } \mu, k + 1\}.$$

We have

$$\begin{aligned} F_z(t) &= Pr\{Z \le t\} = \Gamma_t(\lambda, a-1-m)\Gamma_t(\mu, k+1), \\ u(t) &= F_z'(t) \\ &= f(\lambda, a-1-m;\, t)\Gamma_t(\mu, k+1) \\ &\quad + \Gamma_t(\lambda, a-1-m)\, f(\mu, k+1;\, t) \\ &= u_1(t) + u_2(t), \quad \text{say.} \end{aligned} \tag{4.3.20}$$

Conditioning on the state in which a test unit finds the system, on arrival, we have $Pr\{t \le W_q < t + dt\} = \omega(t)dt$, where

$$\begin{aligned} \omega(t) = \sum_{q-0}^{a-2} f(\lambda, a-1-q;\, t)\, p_{0,q} + \sum_{k=0}^{\infty}\sum_{m=0}^{b-1} f(\mu, k+1;\, t)\, p_{1,kb+m} \\ + \sum_{k=0}^{\infty}\sum_{m=0}^{a-2} u(t)\, p_{1,kb+m}, \quad 0 < t < \infty. \end{aligned} \tag{4.3.21}$$

Substituting the values of $p_{0,q}$ and $p_{1,n}$ [as given in (4.3.13) and (4.3.11), respectively] and the expressions for f and u [given in (4.3.20)] in (4.3.21), we get an explicit expression for $\omega(t)$ as follows.

On simplification, the first term on the RHS of (4.3.21) becomes

$$p_{0,0}\frac{\left[E(a-1, \lambda t) - r^{a-1}\exp(-\lambda t)e\left(a-1, \frac{\lambda t}{r}\right)\right]}{(1-r)}. \tag{4.3.22}$$

The second member on the RHS of (4.3.21) reduces to

$$\begin{aligned} &\left(\frac{\lambda}{\mu}\right) p_{0,0}\left[\sum_{m=a-1}^{b-1} r^m\right]\left[\mu\exp(-\mu t)\sum_{k=0}^{\infty}\frac{(\mu t r^b)^k}{k!}\right] \\ &= \frac{[\lambda p_{0,0}(r^{a-1} - r^b)\ \exp\{-\mu t(1-r^b)\}]}{(1-r)}. \end{aligned}$$

We break up the third term into two parts u_1 and u_2 (corresponding to $u(t) = u_1(t) + u_2(t)$). The part corresponding to u_1 equals

$$\begin{aligned} &\left(\frac{\lambda}{\mu}\right) p_{0,0}\left[\sum_{r=0}^{a-2} r^m f(\lambda, a-1-m;\, t)\right]\int_0^t \left\{\sum_{k=0}^{\infty}(r^b)^k f(\mu, k+1;\, x)\right\} dx \\ &= p_{0,0} r^{a-1}\exp(-\lambda t)\, e\left(a-1, \frac{\lambda t}{r}\right) \times \frac{[1 - \exp\{-\mu(1-r^b)t\}]}{(1-r)}, \end{aligned} \tag{4.3.23}$$

and the part corresponding to u_2 equals

$$\left(\frac{\lambda}{\mu}\right) p_{0,0}\left[\sum_{k=0}^{\infty}(r^b)^k f(\mu, k+1; t)\right]\left[\sum_{m=0}^{a-2} r^m \Gamma_t(\lambda, a-1-m)\right]$$

$$= \lambda p_{0,0}[\exp\{-\mu(1-r^b)t\}]\left[\sum_{m=0}^{a-2} r^m\left\{1 - \sum_{q=0}^{a-2-m}\frac{e^{-\lambda t}(\lambda t)^q}{q!}\right\}\right]$$

$$= \lambda p_{0,0}[\exp\{-\mu(1-r^b)t\}]\left[\frac{1-r^{a-1}}{1-r} - \frac{E(a-1,\lambda t)}{1-r}\right]$$

$$+ \frac{r^{a-1}\exp(-\lambda t)e\left(a-1, \lambda\frac{t}{r}\right)}{1-r}. \tag{4.3.24}$$

Now, adding the expressions on the RHS of (4.3.22) to (4.3.24), we get

$$\omega(t) = \left[\frac{\lambda p_{0,0}}{(1-r)}\right]((1-r^b)\exp\{-\mu(1-r^b)t\}$$

$$+ E(a-1,\lambda t)\{1-\exp[-\mu(1-r^b)t]\}), \quad 0 < t < \infty. \tag{4.3.25}$$

(We may write $\lambda(1-r)/r$ in place of $\mu(1-r^b)$.)

The PDF of the conditional waiting-time distribution given that an arrival has to wait (i.e., an arriving unit who finds the service channel busy or idle with less than $a-1$ waiting) is given by

$$v(t) = \frac{\omega(t)}{[1-p_{0,a-1}]}. \tag{4.3.26}$$

4.3.2.1 Moments of the distribution of W_q

To find the moments, we note that, for positive integral values of k and for $a \geq 2$,

$$J_1(k) = \int_0^{\infty} t^k \exp\{-\mu(1-r^b)t\}dt = \frac{\Gamma(k+1)}{\{\mu(1-r^b)\}^{k+1}} \tag{4.3.27}$$

$$J_2(k) = \int_0^{\infty} t^k E(a-1,\lambda t)dt = \sum_{s=0}^{a-2}\int_0^{\infty}\frac{t^k(\lambda t)^s e^{-\lambda t}dt}{s!}$$

$$= \sum_{s=0}^{a-2}\frac{1}{\lambda}\prod_{i=1}^{k}\left(\frac{s+i}{\lambda}\right) \tag{4.3.28}$$

$$J_3(k) = \int_0^{\infty} t^k E(a-1,\lambda t)\exp\{-\mu(1-r^b)t\}dt$$

$$= \frac{1}{\lambda}\sum_{s=0}^{a-2}\left(\frac{1}{r^{s+1}}\right)\prod_{i=1}^{k}\left(\frac{s+i}{\lambda}\right). \tag{4.3.29}$$

We then put $E(W_q)$ in terms of the above quantities as

$$E(W_q) = \int_0^\infty t\omega(t)\,dt = \left[\frac{\lambda p_{0,0}}{1-r}\right][(1-r^b)J_1(1) + J_2(1) - J_3(1)]$$
$$= \frac{\lambda p_{0,0}}{1-r}\left[\frac{1}{\mu^2(1-r^b)} + \frac{a(a-1)}{2\lambda^2} + \frac{ar^{a+1}(1-r) - r^2(1-r^a)}{\lambda^2(1-r)^2}\right]. \tag{4.3.30}$$

Using (4.3.10), we get

$$\mu^2(1-r^b) = \frac{\lambda^2(1-r)^2}{r^2(1-r^b)}$$

and, thus,

$$E(W_q) = \frac{p_{0,0}}{\lambda(1-r)}\left[\frac{r^2(1-r^b)}{(1-r)^2} + \frac{a(a-1)}{2} + \frac{r^2\{ar^{a-1}(1-r) - (1-r^a)\}}{(1-r)^2}\right]. \tag{4.3.31}$$

Comparing (4.3.15) and (4.3.31), we can at once verify that Little's formula $L_q = \lambda W_q$ holds (for such a bulk-service system also). Since (4.3.27)–(4.3.30) hold for all positive integral values of k, higher moments $E\{W_q^k\}, k = 2, 3, \ldots$ can be obtained easily.

4.3.2.2 Particular case: $M/M(1, b)/1$

We have for $a = 1$,

$$p_{0,0} = \frac{(1-r)}{1-r+\left(\frac{\lambda}{\mu}\right)} = \frac{(1-r)^2}{(1-r)^2 + r(1-r^b)}.$$

$$p_{1,n} = \frac{(1-r)(1-r^b)}{(1-r)^2 + r(1-r^b)} r^{n+1}, \quad n = 0, 1, 2, \ldots$$

The waiting-time density is given by

$$\omega(t) = \left[\frac{\lambda p_{0,0}}{(1-r)}\right][(1-r^b)\exp\{-\mu(1-r^b)t\}], \tag{4.3.32}$$

and the expected waiting time by

$$E(W_q) = \frac{r}{\mu[(1-r)^2 + r(1-r^b)]}. \tag{4.3.33}$$

The conditional waiting-time distribution of only those who wait is exponential with mean $1/\mu\{1-r^b\}$.

When $b = 1$, then $r = \rho$, and the corresponding results for an $M/M/1$ queue follow.

For the model $M/M(a,\infty)/1, r = \lambda/(\lambda+\mu)$; a model with large b has been considered by Sim and Templeton (1983).

Notes:

(1) The unique positive root $r(<1)$ of the equation $h(z) = 0$ is independent of a. Values of r for certain values of b and ρ are given by Cromie and Chaudhry (1976).

(2) Medhi (1979) has given the values of $E(W_q)/b$ and var $(W_q)/b^2$ for certain values of a, b, and ρ, and also their limiting values for large b and for $\theta = a/b = 0(0.1)(1.0)$.

4.3.3 Service batch-size distribution

This has been obtained by Sim and Templeton (1985) for the more general c-server $M/M(a,b)/c$ system.

Let Y be the service batch size for a randomly chosen customer. Then it is a RV that assumes values between a and b. Let k_j be the probability that j customers arrive during an interval T having gamma distribution with parameters $\mu, k+1$. Then

$$\begin{aligned} k_j &= \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^j}{j!} \frac{(\mu)^{k+1}(t)^k e^{-\mu t}}{\Gamma(k+1)}\, dt \\ &= \frac{\lambda^j \mu^{k+1}}{j!k!} \int_0^\infty e^{-(\lambda+\mu)t} t^{(k+j)}\, dt \\ &= \frac{\lambda^j \mu^{k+1}}{j!k!} \frac{(k+j)!}{(\lambda+\mu)^{k+j+1}} \\ &= \binom{k+j}{j} \left(\frac{\lambda}{\lambda+\mu}\right)^j \left(\frac{\mu}{\lambda+\mu}\right)^{k+1}, \quad j = 0,1,2,\ldots \end{aligned} \tag{4.3.34}$$

Consider the four states in which an arriving customer may find the system.

In case he finds the system in state (i) $(0, a-1)$ or (ii) $(0,q)$, with $0 \le q \le a-2$, he will eventually be served in a batch of size exactly a.

In case he finds the system in state (iii) $(1,n)$, with $n = kb+m, a-1 \le m \le b-1$, he will have to wait for the completion of services of $(k+1)$ batches, the duration of which equals T. The size of the service depends on the number of customers who arrive during the interval T. Thus, for $k = 0,1,2,\ldots, a-1 \le m \le b-1$,

$$Pr\{Y = y \mid \text{state}(1, kb+m)\} = \begin{cases} 0, & y \le m \\ k_{y-m-1}, & m < y < b-1 \\ \sum_{i=b-m-1}^{\infty} k_i, & y = b. \end{cases} \tag{4.3.35}$$

In case he finds the system in state (iv), $(1, n)$, $n = kb + m, 0 \le m \le a - 2$, the size of the batch will depend on the number of customers who arrive during the interval T. Thus, for $k = 0, 1, \ldots,\ 0 \le m \le a - 2$,

$$Pr\{Y = y \mid \text{state}(1, kb + m)\} = \begin{cases} \sum_{i=0}^{a-m-1} k_i, & y = a \\ k_{y-m-1}, & a < y \le b - 1 \\ \sum_{i=b-m-1}^{\infty} k_i, & y = b. \end{cases} \quad (4.3.36)$$

It follows that the service batch-size distribution for a *randomly chosen customer* is given by

$$g_a = Pr\{Y = a\} = \sum_{j=0}^{a-1} p_{0,j} + \sum_{k=0}^{\infty} \sum_{m=0}^{a-2} \sum_{i=0}^{a-m-1} k_i\, p_{1,kb+m} \quad (4.3.37)$$

$$g_y = Pr\{Y = y\} = \sum_{k=0}^{\infty} \sum_{m=0}^{y-1} k_{y-m-1}\, p_{1,kb+m}, \quad a < y \le b - 1 \quad (4.3.38)$$

$$g_b = Pr\{Y = b\} = \sum_{k=0}^{\infty} \sum_{m=0}^{b-1} \sum_{i=b-m-1}^{\infty} k_i\, p_{1,kb+m}. \quad (4.3.39)$$

Using the expressions for the state probabilities and simplifying, one can get explicit expressions for g_a, g_y, and g_b (Sim and Templeton, 1985).

Note: Neuts and Nadarajan (1982) consider the same case. The distribution of the size of a randomly chosen batch has not been considered here.

4.4 $M/M(a, b)/1$: Transient-State Distribution

Let us assume that time is reckoned from the instant the server has taken a batch for service, leaving none in the queue—that is, $p_{1,0}(0) = 1$. Let $p^*_{1,n}(s)$, $p^*_{0,q}(s)$ denote, respectively, the LT of $p_{1,n}(t)$ and $p_{0,q}(t)$. Taking LT of Eqs. (4.3.1)–(4.3.4) we get

$$(s + \lambda + \mu)\, p^*_{1,n}(s) = \lambda p^*_{1,n-1}(s) + \mu p^*_{1,n+b}(s), \quad n \ge 1 \quad (4.4.1)$$

$$(s + \lambda + \mu)\, p^*_{1,0}(s) - 1 = \lambda p^*_{0,a-1}(s) + \mu \sum_{r=a}^{b} p^*_{1,r}(s) \quad (4.4.2)$$

$$(s + \lambda)\, p^*_{0,0}(s) = \mu p^*_{1,0}(s) \quad (4.4.3)$$

$$(s + \lambda)\, p^*_{0,q}(s) = \lambda p^*_{0,q-1}(s) + \mu p^*_{1,q}(s), \quad (4.4.4)$$

$$q = 1, 2, \ldots, a - 1 (> 0).$$

and the last equation will not occur for $a = 1$. The equations can now be solved in the same manner as in the case of steady-state solutions. Equation (4.4.1) can be written as

$$h_1(E)\{p^*_{1,n}(s)\} = 0, \quad n = 1, 2, \ldots,$$

where

$$h_1(z) = \mu z^{b+1} - (s + \lambda + \mu)z + \lambda = 0. \tag{4.4.5}$$

Equation (4.4.5) will have only one real zero inside $|z| = 1$, if and only if $\rho = \lambda/b\mu < 1$. Denote this real root by $R \equiv R(s)$ and the other roots by $R_1, \ldots, R_b$. Note that as $s \to 0$, $R \to r$, $R_i \to r_i$. Thus,

$$p^*_{1,n}(s) = AR^n + \sum_{i=1}^{b} A_i R_i^n, \quad n = 0, 1, 2, \ldots$$

Since

$$\sum_n p^*_{1,n}(s) + \sum_q p^*_{0,q}(s) = \frac{1}{s}, \tag{4.4.6}$$

we have $A_1 = 0$ for all i, so that

$$\begin{aligned} p^*_{1,n}(s) &= AR^n, \quad n = 1, 2, \ldots, \\ &= p^*_{1,0}(s)R^n \quad \text{(choosing } A \text{ such that (4.4.1) is satisfied for } n = 1), \\ &= \left(\frac{s + \lambda}{\mu}\right) p^*_{0,0}(s)R^n \quad \text{from (4.4.3).} \end{aligned} \tag{4.4.7}$$

From (4.4.2), using (4.4.7) and simplifying, we get

$$\begin{aligned} p^*_{0,a-1}(s) = {} & \frac{(s + \lambda)(s + \lambda + \mu)}{\lambda\mu} \\ & - \left(\frac{s + \lambda}{\lambda}\right) \frac{R^a(1 - R^{b-a+1})}{1 - R} p^*_{0,0}(s) - \frac{1}{\lambda}. \end{aligned} \tag{4.4.8}$$

Then putting $q = a - 1, \ldots, 1$ we get for $q = a - 2, \ldots 1$,

$$p^*_{o,q}(s) = \left(\frac{\lambda}{s + \lambda}\right)^q \left\{\frac{\lambda}{\lambda - (s + \lambda)R}\right\} \left[1 - \left\{\frac{(s + \lambda)R}{\lambda}\right\}^{q+1}\right] p^*_{0,0}(s). \tag{4.4.9}$$

Finally, using (4.4.6), we get

$$[p^*_{0,0}(s)]^{-1} = \frac{\lambda s}{\mu(1-R)} - \frac{\lambda s(R-R^a)}{(1-R)(\lambda - sR - \lambda R)} + \frac{s(s+\lambda+\mu)}{\mu} - \frac{s(R^a - R^{b+1})}{1-R} + \frac{\lambda^2}{\lambda - sR - \lambda R}\left\{1 - \left(\frac{\lambda}{s+\lambda}\right)^{a-1}\right\}. \tag{4.4.10}$$

This can also be obtained from the relation (4.4.4), using expressions for $p^*_{0,a-1}(s)$, $p^*_{0,a-2}(s)$, $p^*_{1,a-1}(s)$ in terms of $p^*_{0,0}(s)$. Equations (4.4.7)–(4.4.10) give the LT's of the state probabilities.

4.4.1 Steady-state solution

The steady-state probabilities can be obtained easily by taking limits. When steady state exists, we have

$$p_{1,n} = \lim_{t\to\infty} p_{1,n}(t) = \lim_{s\to 0} s p^*_{1,n}(s).$$

Thus, $\lim_{s\to 0} R = r$, where r is the *unique* real root in $(0,1)$ of the equation $h_1(z) = 0$ as $s \to 0$—that is, of $h(z) = 0$ (Eq. (4.3.9)) and r satisfies Eq. (4.3.10). The steady-state probabilities can thus be found, and these agree with those given in Eqs. (4.3.11)–(4.3.14).

4.4.2 Busy-period distribution

The busy-period distribution of the $M/M/(a,b)/1$ system can be obtained in a manner similar to that used for the simple queue $M/M/1$ in Section 3.9.2. We consider a process that avoids the states $0, 1, \ldots, a-1$, immediately after completion of a service. Here a busy period T commences with the start of a service (a batch with a units) and lasts until, for the first time on completion of a service, there are X units ($X = 0, 1, \ldots, a-1$) left in the queue. Let the joint distribution of T and X be given by

$$F_j(t) = Pr\{T \le t, X = j\}, \tag{4.4.11}$$

$$f_j(t)dt = Pr\{t \le T < t + dt, X = j\}, \quad j = 0, 1, \ldots, a-1, \tag{4.4.12}$$

and $f^*_j(s)$ be the LT of $f_j(t)$.

Let us consider the process that avoids the states j, $j = 0, 1, \ldots, a-1$, when the server becomes free. Then

$$f_j(t) = \mu p_{1,j}(t), \quad j = 0, 1, \ldots, a-1, \tag{4.4.13}$$

where $p_{1,j}(t)$ is determined by the following equations:

$$p'_{1,n}(t) = -(\lambda + \mu)\, p_{1,n}(t) + \lambda p_{1,n-1}(t) + \mu p_{1,n+b}(t), \quad n \geq 1 \tag{4.4.14}$$

$$p'_{1,0}(t) = -(\lambda + \mu)\, p_{1,0}(t) + \mu \sum_{r=a}^{b} p_{1,r}(t). \tag{4.4.15}$$

From the definition of busy period, we have $p_{1,0}(0) = 1$. Taking the LT of (4.4.14) and (4.4.15) and using $p_{1,0}(0) = 1$, we get

$$(s + \lambda + \mu)\, p^*_{1,n}(s) = \lambda p^*_{1,n-1}(s) + \mu p^*_{1,n+b}(s), \quad n \geq 1 \tag{4.4.16}$$

$$(s + \lambda + \mu)\, p^*_{1,0}(s) = 1 + \mu \sum_{r=a}^{b} p^*_{1,r}(s). \tag{4.4.17}$$

From (4.4.16) we get as before

$$p^*_{1,n}(s) = R^n p^*_{1,0}(s) \tag{4.4.18}$$

and using it, we get from (4.4.17)

$$p^*_{1,0}(s) = \frac{1 - R}{\mu - \mu R^a + s}. \tag{4.4.19}$$

Finally

$$p^*_{1,n}(s) = \frac{R^n(1 - R)}{s + \mu - \mu R^a}, \quad n \geq 0. \tag{4.4.20}$$

Thus, for $0 \leq j \leq a - 1$

$$f^*_j(s) = \frac{\mu(1 - R)\, R^j}{s + \mu - \mu R^a} \tag{4.4.21}$$

and the LT $b^*(s)$ of the busy-period PDF $b(t)$ of T is given by

$$\begin{aligned} b^*(s) = \sum_{j=0}^{a-1} f^*_j(s) &= \frac{\mu(1 - R)}{s + \mu - \mu R^a} \sum_{j=0}^{a-1} R^j \\ &= \frac{\mu(1 - R^a)}{s + \mu - \mu R^a}. \end{aligned} \tag{4.4.22}$$

Inverting (4.4.22), one gets the PDF $b(t)$ of T

$$b(t) = \sum_{j=0}^{a-1} f_j(t). \tag{4.4.23}$$

We have

$$b^*(0) = \sum_{j=0}^{a-1} f_j^*(0)$$

$$= \int_0^\infty b(t)dt = 1. \tag{4.4.24}$$

Now $f_j^*(0)$ is the probability that the busy period terminates, leaving j $(0 \leq j \leq a-1)$ in the queue. The corresponding idle periods are distributed as gamma with parameters λ, j. We have

$$f_j^*(0) = \lim_{s\to 0} f_j^*(s)$$

$$= \frac{(1-r)r^j}{(1-r^a)}, \quad j = 0, 1, \ldots, a-1, \tag{4.4.25}$$

where r is the unique root in $(0, 1)$ of (4.3.9) ($r = \lim R$ as $s \to 0$).

4.4.2.1 Mean and variance of the busy period

Moments of T can be obtained by differentiating $b^*(s)$. From (4.4.22).

$$b^*(s) = 1 - \frac{s}{s + \mu - \mu R^a}. \tag{4.4.26}$$

We have

$$E(T) = -\frac{d}{ds} b^*(s)|_{s=0}$$

$$= \frac{(s+\mu-\mu R^a) - s(1-\mu a R^{a-1}R')}{(s+\mu-\mu R^a)^2}\bigg|_{s=0}$$

$$= \frac{\mu(1-r^a)}{[\mu(1-r^a)]^2} = \frac{1}{\mu(1-r^a)}, \tag{4.4.27}$$

where $R' = (d/ds)R = R'(s)$. As R satisfies (4.4.5), we have

$$\mu R^{b+1} - (s + \lambda + \mu)R + \lambda = 0.$$

Differentiating, we get

$$\mu(b+1)R^b R' - R - (s+\lambda+\mu)R' = 0 \quad \text{or}$$

$$R' = \frac{R}{\mu(b+1)R^b - (s+\lambda+\mu)},$$

and at $s = 0$, we get

$$R_0' = \frac{r}{\mu(b+1)r^b - (\lambda+\mu)} \tag{4.4.28}$$

$$= \frac{1}{\mu} \frac{r}{(b+1)r^b - 1 - \frac{r(1-r^b)}{1-r}}$$

$$= \frac{1}{\mu} \frac{(1-r)r}{br^b(1-r) - (1-r^b)}, \quad \text{and} \tag{4.4.28a}$$

$$E(T^2) = \frac{d^2}{ds^2} b^*(s)|_{s=0}$$

$$= \frac{2[1 - a\mu r^{a-1} R_0']}{[\mu(1-ra)]^2}, \tag{4.4.29}$$

where R_0' is given by (4.4.28). We have

$$\text{var}(T) = \frac{1}{[\mu(1-r^a)]^2} - \left(\frac{2a\mu r^{a-1}}{[\mu(1-ra)]^2}\right)\left(\frac{1}{\mu}\right)\left(\frac{(1-r)r}{br^b(1-r) - (1-r^b)}\right) \tag{4.4.30}$$

$$= E(T)^2 + \frac{2ar^a(1-r)}{[\mu(1-r^a)]^2[(1-r^b) - br^b(1-r)]}$$

$$= \frac{2ar^a(1-r) + (1-r^b) - br^b(1-r)}{[\mu(1-r^a)]^2[(1-r^b) - br^b(1-r)]}. \tag{4.4.30a}$$

4.4.2.2 Particular cases

***M/M*(1, *b*)/1**

When $a = 1$, the busy period always terminates with none in the queue (i.e., termination of a busy period leaves the system empty), and in this case the busy and idle periods alternate. We can then use the result

$$\frac{E(T)}{E(I)} = \frac{1 - p_{0,0}}{p_{0,0}}$$

to find the mean busy period $E(T)$. We have $E(I) = 1/\lambda$, and when $a = 1$.

$$p_{0,0} = \frac{(1-r)^2}{1 - r + r^2 - r^{b+1}},$$

so that

$$E(T) = \frac{1}{\lambda} \frac{1 - p_{0,0}}{p_{0,0}} = \frac{1}{\lambda} \frac{r(1-r^b)}{(1-r)^2} = \frac{1}{\mu(1-r)}.$$

We get the same result as given by (4.4.27) with $a = 1$. Putting $a = 1$ in (4.4.30)

$$\text{var}(T) = \left[\frac{1}{\mu(1-r)}\right]^2 + \frac{2r}{\mu^2(1-r)[(1-r^b) - br^b(1-r)]}.$$

M/M(k, k)/1

Putting $a = b = k$ in (4.4.27) and (4.4.30), we get

$$E(T) = \frac{1}{\mu(1-r^a)} \quad \text{and}$$

$$\text{var}(T) = \frac{kr^k(1-r) + (1-r^k)}{[\mu(1-r^k)]^2[(1-r^k) - kr^k(1-r)]},$$

where r is the unique root in $(0, 1)$ of

$$\mu z^{k+1} - (\lambda + \mu)z + \lambda = 0.$$

M/M/1

Putting $k = 1$ (and noting that $r = \rho$), and get

$$E(T) = \frac{1}{\mu(1-\rho)} \quad \text{and}$$

$$\text{var}(T) = \frac{1-\rho^2}{[\mu(1-\rho)]^2[(1-\rho) - \rho(1-\rho)]}$$

$$= \frac{1+\rho}{\mu^2(1-\rho)^3}.$$

Note: Selim (1997) examines an $M/M(1, N)/1/N$ bulk-service queue with finite capacity waiting room; a traffic engineering problem for which this could serve as a model is cited. In addition to giving an explicit transient solution, he studies an optimal control problem of the bulk-service queue.

More general models with finite waiting space are considered (see Chapter 6).

4.5 Two-Server Model: *M/M(a, b)/2*

Consider that there are two parallel (homogeneous) servers and service is under a general bulk-service rule. When both the servers are free, each is equally likely to take a batch for service. Let

$p_{r,n}(t) = Pr\{$at time t, r channels are busy, and there are n waiting in the queue$\}$;

$p_{r,n} = \lim_{t \to 0} p_{r,n}(t)$

$\rho = \dfrac{\lambda}{2\mu b}$ = the traffic intensity.

We have the following differential-difference equations governing $p_{r,n}(t)$:

$$p'_{2,n}(t) = -(\lambda + 2\mu)\, p_{2,n}(t) + \lambda p_{2,n-1}(t) + 2\mu p_{2,n+b}(t), \quad n \geq 1 \tag{4.5.1}$$

$$p'_{2,0}(t) = -(\lambda + 2\mu)\, p_{2,0}(t) + \lambda p_{1,a-1}(t) + 2\mu \sum_{k=a}^{b} p_{2,k}(t) \tag{4.5.2}$$

$$p'_{1,0}(t) = -(\lambda + \mu)\, p_{1,0}(t) + \lambda p_{0,a-1}(t) + 2\mu p_{2,0}(t) \tag{4.5.3}$$

$$p'_{1,q}(t) = -(\lambda + \mu)\, p_{1,q}(t) + \lambda p_{1,q-1}(t) + 2\mu p_{2,q}(t), \tag{4.5.4}$$

$$1 \leq q \leq a - 1$$

$$p'_{0,q}(t) = -\lambda p_{0,q}(t) + \lambda p_{0,q-1}(t) + \mu p_{1,q}(t), \quad 1 \leq q \leq a - 1 \tag{4.5.5}$$

$$p'_{0,0}(t) = -\lambda p_{0,0}(t) + \mu p_{1,0}(t). \tag{4.5.6}$$

It is to be noted that (4.5.4) and (4.5.5) will not occur when $a = 1$. Assume that steady-state solutions exist. Then $p_{r,n}$ will be the solutions of the following equations:

$$0 = -(\lambda + 2\mu)\, p_{2,n} + \lambda p_{2,n-1} + 2\mu p_{2,n+b}, \quad n \geq 1 \tag{4.5.7}$$

$$0 = -(\lambda + 2\mu)\, p_{2,0} + \lambda p_{1,a-1} + 2\mu \sum_{k=a}^{b} p_{2,k} \tag{4.5.8}$$

$$0 = -(\lambda + \mu)\, p_{1,0} + \lambda p_{0,a-1} + 2\mu p_{2,0} \tag{4.5.9}$$

$$0 = -(\lambda + \mu)\, p_{1,q} + \lambda p_{1,q-1} + 2\mu p_{2,q}, \quad 1 \leq q \leq a - 1 \tag{4.5.10}$$

$$0 = -\lambda p_{0,q} + \lambda p_{0,q-1} + \mu p_{1,q}, \quad 1 \leq q \leq a - 1 \tag{4.5.11}$$

$$0 = -\lambda p_{0,0} + \lambda p_{1,0}. \tag{4.5.12}$$

As before, (4.5.10) and (4.5.11) will not occur for $a = 1$, and either (4.5.9) or (4.5.12) may be used. These equations can be solved very much the same way. Equation (4.5.7) can be written as

$$g(E)[p_{2,n}] = 0, \quad n = 0, 1, 2, \ldots,$$

$$\text{where} \quad g(z) \equiv 2\mu z^{b+1} - (\lambda + 2\mu)z + \lambda = 0. \tag{4.5.13}$$

Note that (4.5.13) is obtained by writing 2μ in place of μ in (4.3.9). $g(z)$ has a unique real root in (0, 1) if and only if $\rho = \lambda/2b\mu < 1$. Denote this real root by r; r satisfies

$$p = b\rho = \frac{\lambda}{2\mu} = \frac{r(1 - r^b)}{1 - r} \tag{4.5.14}$$

(the same as (4.3.11), but with μ replaced by 2μ). Thus, it follows that

$$p_{2,n} = p_{2,0} r^n, \quad n \geq 0. \tag{4.5.15}$$

Substituting this value of $p_{2,n}$ in (4.5.8), one gets $p_{1,a-1}$; then from (4.5.10) $p_{1,a-2}, p_{1,a-3}, \ldots, p_{1,0}$. Using the value of $p_{1,0}$, one gets $p_{0,a-1}$ from (4.5.9) and from (4.5.11), $p_{0,a-2} \ldots, p_{0,0}$, and finally $p_{1,0}$ from (4.5.12). Thus, all the probabilities can be found in terms of $p_{2,0}$. Using the normalizing condition

$$\sum_{n=0}^{\infty} p_{2,n} + \sum_{q=0}^{a-1} p_{1,q} + \sum_{q=0}^{a-1} p_{0,q} = 1,$$

one gets $p_{2,0}$. Thus, all the probabilities can be completely obtained.

Note: $p_{2,0}$ can also be obtained by using (4.4.11) with $q = a - 1$ and expressing $p_{0,a-1}$, $p_{0,a-2}$, and $p_{1,a-1}$ all in terms of $p_{2,0}$.

See Medhi and Borthakur (1972) for further results.
Chaudhry *et al.* (1987) discuss a heterogeneous server model.

4.5.1 Particular case: $M/M(1, b)/2$

Two-server Markovian queue under usual bulk service.

Using (4.5.15) in (4.5.8) with $a = 1$, we get

$$\begin{aligned} 0 &= -(\lambda + 2\mu) p_{2,0} + \lambda p_{1,0} + 2\mu \sum_{k=1}^{b} p_{2,k} \\ &= -(\lambda + 2\mu) p_{2,0} + \lambda p_{1,0} + 2\mu p_{2,0} \frac{r(1 - r^b)}{1 - r} \end{aligned}$$

and using (4.5.14), we get

$$\begin{aligned} p_{1,0} &= \frac{1}{\lambda}\left[(\lambda + 2\mu) - 2\mu \frac{\lambda}{2\mu}\right] p_{2,0} \\ &= \frac{2\mu}{\lambda} p_{2,0}. \end{aligned}$$

Finally, from (4.5.12) (or from (4.5.9)) we get

$$p_{0,0} = \frac{1}{2}\left(\frac{2\mu}{\lambda}\right)^2 p_{2,0}.$$

Writing in terms of $p_{0,0}$, and $p = \lambda/2\mu$

$$p_{2,n} = p_{2,0} r^n = 2p^2 p_{0,0} r^n, \quad n \geq 0 \tag{4.5.15a}$$

$$p_{1,0} = 2\left(\frac{\lambda}{2\lambda}\right) p_{0,0} = 2p p_{0,0}. \tag{4.5.16}$$

Using $\sum_{n=0}^{\infty} p_{2,n} + p_{1,0} + p_{0,0} = 1$, we get

$$p_{0,0} = \left[1 + 2p + \frac{2p^2}{1-r}\right]^{-1} = \frac{(1-r)}{2p^2 + (2p+1)(1-r)}. \quad (4.5.17)$$

The expected number in the queue $E(Q)$ is given by

$$E(Q) = \sum_{n=0}^{\infty} n p_{2,n}$$

$$= p_{2,0} \sum_{n=0}^{\infty} n r^n = p_{2,0}\left[\frac{r}{(1-r)^2}\right] \quad (4.5.18)$$

$$= \frac{2p^2 r}{(1-r)[2p^2 + (2p+1)(1-r)]}. \quad (4.5.18a)$$

When $b = 1, r = \rho = \lambda/2\mu = p$; we have for an $M/M/2$ queue:

$$p_{0,0} = \frac{1-\rho}{2\rho^2 + (2\rho+1)(1-\rho)} = \frac{1-\rho}{1+\rho},$$

$$p_{2,n} = \frac{(2\rho^{n+2})(1-\rho)}{1+\rho},$$

$$p_{1,0} = \frac{(2\rho)(1-\rho)}{1+\rho}, \quad \text{and}$$

$$E(Q) = \frac{2\rho^3}{(1-\rho)(1+\rho)} = \frac{2\rho^3}{1-\rho^2}. \quad (4.5.19)$$

Note that with our earlier notation for $M/M/2$

$$p_0 = p_{0,0}, \quad p_{1,0} = p_1,$$
$$p_{2,n} = p_{n+2}, \quad n \geq 0,$$

where $p_m = \text{Prob}\{\text{number in the system is } m\}$ in the usual $M/M/2$ notation.

4.6 The $M/M(1, b)/c$ Model

Ghare (1968) considered this model. We discuss it with some modification of the procedures adopted by him. Let

$$p_{r,n}(t) = Pr\{r \text{ channels are busy, } n \text{ units in queue at time } t\}.$$

Let $p^*_{r,n}(s)$ be the LT of $P_{r,n}(t)$ and $p_{r,n}$ be the corresponding steady-state probabilities. The equations involving $p_{r,n}$ satisfy the following equations

$$p'_{c,n}(t) = -(\lambda + c\mu)\,p_{c,n}(t) + \lambda p_{c,n-1}(t) + c\mu p_{c,n+b}(t), \quad n \geq 0 \tag{4.6.1}$$

$$p'_{c,0}(t) = -(\lambda + c\mu)\,p_{c,0}(t) + \lambda p_{c-1,0}(t) + c\mu \sum_{k=1}^{b} p_{c,k}(t) \tag{4.6.2}$$

$$p'_{r,0}(t) = -(\lambda + r\mu)\,p_{r,0}(t) + \lambda p_{r-1,0}(t) + (r+1)\mu p_{r+1,0}(t), \quad 1 \leq r \leq c-1 \tag{4.6.3}$$

$$p'_{0,0}(t) = -\lambda p_{0,0}(t) + \mu p_{1,0}(t). \tag{4.6.4}$$

Let the initial condition be $p_{c,0}(0) = 1$. That is, that time is reckoned from the instant that all the channels become busy, leaving none in the queue. Taking LTs, we get

$$(s + \lambda + c\mu)\,p^*_{c,n}(s) = \lambda p^*_{c,n-1}(s) + c\mu p^*_{c,n+b}(s), \quad n > 0 \tag{4.6.5}$$

$$(s + \lambda + c\mu)\,p^*_{c,0}(s) = 1 + \lambda p^*_{c-1,0}(s) + c\mu \sum_{k=1}^{b} p^*_{c,k}(s) \tag{4.6.6}$$

$$(s + \lambda + r\mu)\,p^*_{r,0}(s) = \lambda p^*_{r-1,0}(s) + (r+1)\mu p^*_{r+1,0}(s), \quad 1 \leq r \leq c-1 \tag{4.6.7}$$

$$(s + \lambda)\,p^*_{0,0}(s) = \mu p^*_{1,0}(s). \tag{4.6.8}$$

Consider the equation

$$h(z) \equiv \lambda - (\lambda + s + c\mu)z + c\mu z^{b+1} = 0. \tag{4.6.9}$$

The equation has a unique root $R = R(s)$ in $(0, 1)$. Then from (4.6.5) we get, as before

$$\begin{aligned} p^*_{c,n}(s) &= AR^n \\ &= p^*_{c,0}(s)\,R^n, \quad n \geq 0. \end{aligned} \tag{4.6.10}$$

Then using (4.6.10), we can find $p^*_{c-1,0}(s)$ from (4.6.6) and then recursively $p^*_{r,0}(s)$, $r = 1, 2, \ldots, c-2$, from (4.6.7). Thus, all the probabilities can be obtained in terms of $p^*_{0,0}(s)$, and this can be evaluated as before. Instead, the general solution of the set of equations (4.6.7) and (4.6.8) can be written (see our treatment of $M/M/c$ queue in Section 3.10 by the method of Jackson and Henderson (1966)) as

$$p^*_{r,0}(0) = p^*_{0,0}(s) \sum_{j=0}^{r} \frac{\left(\frac{\lambda}{\mu}\right)^{r-j} \Gamma\left(j + \frac{s}{\mu}\right)}{\Gamma\left(\frac{s}{\mu}\right) j!(r-j)!} \tag{4.6.11}$$

$$= p^*_{0,0}(s)\phi_r(s), \quad 0 \leq r \leq c, \tag{4.6.11a}$$

where

$$\phi_r(s) = \left[\frac{\left(\frac{\lambda}{\mu}\right)^r}{r!}\right] {}_2F_0\left(-r, \frac{s}{\mu}; -; -\frac{\mu}{\lambda}\right), \tag{4.6.12}$$

and ${}_2F_0$ is a confluent hypergeometric series (see Abramowitz and Stegun (1968)). Using (4.6.11a) and (4.6.6), we get

$$(s + \lambda + c\mu)p^*_{c,0}(s) = 1 + \lambda p^*_{0,0}(s)\phi_{c-1}(s) + c\mu p^*_{c,0}(s)\frac{R(1 - R^b)}{1 - R} \quad \text{or}$$

$$\left[(s + \lambda + c\mu) - \frac{c\mu R(1 - R^b)}{1 - R}\right] p^*_{c,0}(s) = 1 + \lambda p^*_{0,0}(s)\phi_{c-1}(s).$$

Using the fact that R satisfies (4.6.9)—that is, the relation

$$\lambda - (s + \lambda + s + c\mu)R + c\mu R^{b+1} = 0,$$

we get

$$(s + \lambda + c\mu) - \frac{c\mu R(1 - R^b)}{1 - R} = \frac{s + c\mu(1 - R)}{1 - R}.$$

Thus,

$$p^*_{c,0}(s) = \frac{1 - R}{s + c\mu(1 - R)}[1 + \lambda p^*_{0,0}(s)\phi_{c-1}(s)]. \tag{4.6.13}$$

Thus, we get $p^*_{c,n}(s)$, $p^*_{r,0}(s)$ $(r = 1, \ldots, c - 1)$, all in terms of $p^*_{0,0}(s)$. Using the relation

$$\sum_{n=0}^{\infty} p^*_{c,n}(s) + \sum_{r=1}^{c-1} p^*_{r,0}(s) + p^*_{0,0}(s) = \frac{1}{s}, \tag{4.6.14}$$

we get

$$p^*_{c,0}(s)\left(\frac{1}{1 - R}\right) + \sum_{r=1}^{c-1} \phi_r(s)\, p^*_{0,0}(s) + p^*_{0,0}(s) = \frac{1}{s}.$$

Thus, we have

$$p^*_{0,0}(s)\left[1 + \sum_{r=1}^{c-1} \phi_r(s) + \frac{\lambda\phi_{c-1}(s)}{s + c\mu(1 - R)}\right] = \frac{1}{s} - \frac{1}{s + c\mu(1 - R)}$$

$$= \frac{c\mu(1 - R)}{s\{s + c\mu(1 - R)\}}$$

and finally,

$$p_{0,0}^*(s) = \frac{c\mu(1-R)}{s}\left[\lambda\phi_{c-1}(s)\{s + c\mu(1-R)\}\sum_{r=0}^{c-1}\phi_r(s)\right]^{-1} \tag{4.6.15}$$

noting that $\phi_0(s) = 1$.

4.6.1 Steady-state results $M/M(1, b)/c$

The root of $h(z) \equiv 0$ for $s = 0$ is real and unique and will be in $(0, 1)$ IFF $\frac{\lambda}{bc\mu} < 1$—that is, $\frac{\lambda}{c\mu} < b$. Denote this root by

$$\lim_{s\to 0} R = r; \text{ we get}$$

$$\begin{aligned}\lim_{s\to 0}\phi_m(s) &= \frac{\left(\frac{\lambda}{\mu}\right)^m}{m!} + \lim_{s\to 0}\sum_{j=1}^{m}\frac{\left(\frac{\lambda}{\mu}\right)^{m-j}}{j!(m-j)!}\left[\left(\frac{s}{\mu}\right)\left(\frac{s}{\mu}+1\right)\cdots\left(\frac{s}{\mu}+j-1\right)\right] \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^m}{m!},\end{aligned} \tag{4.6.16}$$

since

$$\frac{\Gamma(a+n)}{\Gamma(a)} = \begin{cases} 1, & \text{for } n = 0, \quad a = 0 \\ a(a+1)\cdots(a+n-1), & n = 1, 2, 3, \ldots. \end{cases}$$

Thus,

$$\begin{aligned} p_{0,0} &= \lim_{s\to 0} s p_{0,0}^*(s) = c\mu(1-r)\left[\lambda\frac{\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!} + c\mu(1-r)\sum_{m=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^m}{m!}\right]^{-1} \\ &= \left[\frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(1-r)} + \sum_{m=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^m}{m!}\right]^{-1} \end{aligned} \tag{4.6.17}$$

$$\begin{aligned} p_{m,0} &= \lim_{s\to 0} s p_{m,0}^*(s) = \lim_{s\to 0}[\{s p_{0,0}^*(s)\}\{\phi_m(s)\}] \\ &= p_{0,0}\frac{\left(\frac{\lambda}{\mu}\right)^m}{m!}, \quad m = 1, 2, \ldots, c-1 \end{aligned} \tag{4.6.18}$$

$$\begin{aligned} p_{c,n} &= \lim_{s\to 0} s p_{c,n}^*(s) = \lim_{s\to 0} s p_{c,0}^*(s) R^n \\ &= r^n\left[\frac{\lambda}{c\mu}\cdot\frac{\left(\frac{\lambda}{\mu}\right)^{c-1}}{(c-1)!}p_{0,0}\right] \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}r^n p_{0,0}, \quad n \geq 0. \end{aligned} \tag{4.6.19}$$

We have

$$E(Q) = \frac{(\lambda/\mu)^c}{c!} \frac{r}{(1-r)^2} p_{0,0} \tag{4.6.20}$$

4.6.1.1 Particular cases

(1) By putting $c = 2$, we obtain the results for $M/M(1, b)/2$.

(2) $M/M/c$: Putting $b = 1$, we get $r = \lambda/c\mu = \rho$, so that

$$p_0 = p_{0,0} = \left[\sum_{m=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^m}{m!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!(1-\rho)} \right]^{-1}$$

$$p_n = p_{n,0} = \left[\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \right] p_{0,0}, \quad n \leq c$$

$$p_n = p_{c,n-c} = \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} \rho^{n-c} p_{0,0}, \quad n \geq c,$$

where $p_n = Pr\{\text{number in system is } n\}$ in an $M/M/c$ queue.

M/M(a, b)/c. This model has been investigated by Neuts and Nadarjan (1982) and Sim and Templeton (1985). We mention here a simple result relating to its busy period.

Suppose that the busy period T starts at the instant when all the servers become busy and lasts until the instant when any one of the servers becomes free for the first time. Then the LT of the busy period can be obtained from that of $M/M(a, b)/1$ by replacing μ by $c\mu$. Thus, the LST of the busy period T is given by

$$b^*(s) = \frac{c\mu(1 - R^a)}{s + c\mu(1 - R^a)},$$

and the probability that the busy period terminates with j in the queue is

$$f_j^*(0) = \frac{(1-r)r^j}{(1-r^a)}, \quad 0 \leq j \leq a-1.$$

The mean busy period is $E(T) = 1/c\mu(1 - r^a)$.

M/M(α, ∞)/N queue has been discussed by Cosmetatos (1983a).

Problems and Complements

4.1. The system $M/E_k/1$: transient behavior. Define

$$P(s,t) = \sum_{n=0}^{\infty} p_n(t)s^n, \qquad P^*(s,\alpha) = \sum_{n=0}^{\infty} p_n^*(\alpha)s^n,$$

where $p_n^*(\alpha)$ is the LT of $p_n(t)$. Assume that $p_0(0) = 1$. Show that

$$P^*(s,\alpha) = \frac{s - k\mu(1-s)\,p_0^*(\alpha)}{s(\alpha + \lambda + k\mu) - k\mu - \lambda s^{k+1}}, \quad |s| < 1,$$

and that

$$\lim_{\alpha \to 0} p_0^*(\alpha) = \frac{c}{k\mu(1-c)},$$

where $c \equiv c(s,\alpha)$ is the only zero of $(\alpha + \lambda + k\mu)s - k\mu - \lambda s^{k+1}$ inside the unit circle $|s| = 1$. Noting that

$$\lim_{\alpha \to 0} \alpha p_n^*(\alpha) = \lim_{t \to \infty} p_n(t) = p_n \quad \text{and}$$

$$\lim_{\alpha \to 0} \alpha P^*(s,\alpha) = \lim_{t \to \infty} P(s,t) = P(s) = \sum_{n=0}^{\infty} p_n s^n,$$

derive the expression for the PGF $P(s)$ [in the form given in (4.1.6)].

4.2. The system $E_k/M/1$

Write down the Chapman-Kolmogorov equations and obtain an expression for $P^*(s,\alpha)$.

Obtain the relevant equations for finding the distribution of the busy period T (the interval from the arrival of a customer in an empty system to the first subsequent instant when the system becomes empty) of an $E_k/M/1$ system. Show the LT $f^*(s)$ of the PDF of the busy period T is given by

$$f^*(s) = \frac{\mu(\alpha - 1)}{\alpha(s + \mu) - \mu}$$

where $\alpha \equiv \alpha(s)$ is the root lying outside $|z| = 1$ of the equation

$$g(z,s) \equiv z - \left[\frac{z(s + \mu + k\lambda) - \mu}{k\lambda z}\right]^k = 0.$$

Show that the mean busy period is given by

$$E(T) = \frac{\alpha_0}{\mu(\alpha_0 - 1)}$$

where α_0 is the unique root outside $|z| = 1$ of the equation $g(z,0) = 0$.

4.3. The System $M^X/M/1$

(a) Find the variance of the number in the system for the $M^X/M/1$ system.

(b) Find the mean and variance of the number in the system for the $M^X/M/1$ system where X has a geometric distribution.

(c) Find the mean and variance of the number in the system for the $M^r/M/1$ system (where r is a fixed positive integer).

4.4. $M^X/E_k/1$ system (Restrepo, 1965).
Let the state of the system be denoted by $(n, s), n = 0, 1, 2, \ldots, s = 1, 2, \ldots, k$ where n is the number of customers in the system and s is the number of phases that remain to be completed by the person in service, if any. Let $p_{n,s}$ be the steady-state probability that the system is in state $(n, s), n = 1, 2, \ldots; s = 1, \ldots, k$; for $n = 0, p_{0,0} = p_0$. Writing $\theta = \lambda/k\mu$, show that the system of equations can be obtained as follows:

$$0 = p_{1,1} - \theta p_0 \tag{4.c.1}$$

$$0 = p_{1,s+1} - (1+\theta)\, p_{1,s} \tag{4.c.2}$$

$$0 = p_{2,1} - (1+\theta)\, p_{1,k} + \theta a_1 p_0 \tag{4.c.3}$$

$$0 = p_{n,s+1} - (1+\theta)\, p_{n,s} + \theta \sum_{m=1}^{n-1} a_m p_{n-m,s} \tag{4.c.4}$$

$$0 = p_{n+1,1} - (1+\theta)\, p_{n,k} + \theta a_n p_0 + \theta \sum_{m=1}^{n-1} a_m p_{n-m,k} \tag{4.c.5}$$

with (4.c.2), (4.c.4), (4.c.5) restricted to $s < k$ and $n > 1$. How would you proceed to solve the preceding equations in terms of p_0 recursively starting from Eq. (4.c.1)? [$A(z)$ is the PGF of X, and $a_k = Pr\{X = k\}$.]

4.5. For a continuation of Problem 4.4, show that the generating function

$$F(z) = p_0 + \sum_{n=1}^{\infty} \sum_{s=1}^{k} p_{n,s} z^n \tag{4.c.6}$$

is given by the expression

$$F(z) = \frac{p_0(1-z)}{1 - z\{1 + \theta - \theta A(z)\}^k}. \tag{4.c.7}$$

Show that

$$p_0 = 1 - \frac{\lambda \bar{a}}{\mu} = 1 - \rho, \quad (E(X) = \bar{a}, \sigma = \mathrm{var}(X))$$

$$E\{N\} = \rho + E(Q)$$

$$= \rho + \frac{k+1}{2k\mu}\left\{\frac{(\lambda\bar{a})^2}{\mu - \lambda\bar{a}}\right\} + \frac{\lambda}{2(\mu - \lambda\bar{a})}\{\sigma^2 + \bar{a}^2 - \bar{a}\}. \tag{4.c.8}$$

Deduce that for $E_k \equiv E_1 \equiv M$—that is, for $M^X/M/1$

$$E\{N\} = \frac{\lambda(\sigma^2 + \bar{a}^2 + \bar{a})}{2(\mu - \lambda\bar{a})}$$

and for $M^X/D/1$,

$$E\{N\} = \frac{\lambda(\sigma^2 + \bar{a}^2 + \bar{a})}{2(\mu - \lambda\bar{a})} - \frac{(\lambda\bar{a})^2}{2\mu(\mu - \lambda\bar{a})}$$

(Restrepo, 1965).

4.6. Multiple Poisson bulk arrival (MPBA) system (Jensen *et al.* 1977). Consider that groups of size j arrive according to a Poisson process with intensity λ_j, $\{N_j(t), t \geq 0\}$, $j = 1, 2, \ldots, m$, and that these processes are mutually independent. Note that the group size is restricted to m. Then $M(t) = \sum_{j=1}^{m} N_j(t)$ denotes the number of groups that arrive in $(0, t)$, while the compound Poisson process $\{N(t), t \geq 0\}$, where

$$N(t) = \sum_{j=1}^{m} jN_j(t)$$

gives the number of arrivals in $(0, t)$. The mean arrival rate of customers is $\sum_{j=1}^{m} j\lambda_j$. Then the arrival distribution X is given by

$$a_j = Pr(X = j) = \frac{\lambda_j}{\lambda}\delta(m - j), \quad j = 1, 2, \ldots, m,$$

where

$$\begin{aligned} \delta(x) &= 1 \quad \text{for } x \geq 0 \\ &= 0 \quad \text{for } x < 0. \end{aligned}$$

For such a compound Poisson arrival process and exponential service, show that the steady-state probabilities p_n of the distribution of number in the system satisfy

$$p_{k+1} = \sum_{j=1}^{m} \rho_j p_{k+1-j}, \quad k = 0, 1, 2, \ldots,$$

where

$$\rho_j = \sum_{k=j}^{m} \frac{\lambda_k}{\mu}, \quad j = 1, \ldots, m, \quad p_r = 0, \quad r < 0.$$

(This relation is deducible from (4.2.3a) and (4.2.3b).) Further show that PGF $P(s) = \sum_{n=0}^{\infty} p_n s^n$ is given by

$$P(s) = \frac{p_0}{1 - \sum_{j=1}^{m} p_j s^j},$$

where

$$p_0 = 1 - \sum_{j=1}^{m} p_j = 1 - \sum_j \frac{\lambda_j}{\mu} = 1 - \rho.$$

4.7. $M^X/M/\infty$ queue. Show that given $N(0) = i$, the correlation coefficient ρ between $N(t)$ and $N(0)$ equals $e^{-\mu t}$ and is independent of the arrival rate λ and of the batch size X. As $t \to \infty, \rho \to 0$, as is expected (Reynolds, 1968).

4.8. $M^r/M/c$ (Kabak, 1968) queue with rates λ and $\mu = 1$
Consider (i) the delay system $M^r/M/c/\infty$, where customers are allowed to wait, and (ii) the loss system $M^r/M/c/c$, where customers are lost when all the channels are busy. If $\{p_i\}$ is the steady-state probability distribution of the system size, show that p_j satisfies the following equations.

For the *delay* system,

$$p_i = \frac{\lambda}{i} \sum_{j=\max(0,i-r)}^{i-1} p_j, \quad 1 \le i \le c-1,$$

$$= \frac{\lambda}{c} \sum_{j=\max(0,i-r)}^{i-1} p_j, \quad i \ge c, \quad \text{and}$$

$$\sum_{i=0}^{\infty} p_i = 1.$$

For the *loss* system,

$$p_i = \frac{\lambda}{i} \sum_{j=0}^{i-1} p_j, \quad 1 \le i \le c, \quad i \le r$$

$$= \frac{\lambda}{i} \sum_{j=i-r}^{i-1} p_j, \quad 1 \le i \le c, \quad i \ge r, \quad \text{and}$$

$$\sum_{i=0}^{c} p_i = 1.$$

4.9. For a continuation of Problem 4.8, consider waiting-time distribution in the delay system. Show that the distribution of waiting time W in the

queue

$$P(W > t) = \left(\frac{1}{r}\right) \sum_{d=1}^{\infty} \sum_{i=\max(0,d+c-r)}^{d+c-1} p_i F,$$

where

$$F = \frac{\Gamma(d, ct)}{\Gamma(d)}, \quad \text{and}$$

$$\Gamma(a, x) = \int_x^{\infty} e^{-x} x^{a-1} dx$$

(Kabak, 1968).

4.10. Model $M/M(1, b; \mu_k)/1$, with mean service time $1/\mu_k$ for a batch of k. Distribution of queue size and number in the batch being served in Poisson queue with usual bulk service has been considered by Cosmetatos (1983c). Show that the steady-state difference equations for $p(n; k)$, where $p(n; k)$ denotes the steady-state probability of n customers being in the queue and k being served ($0 \leq k \leq b$), satisfy the following equations:

$$0 = -p(0; 0) + \sum_{k=1}^{b} \mu_k p(0; k)$$

$$0 = -(\lambda + \mu_1) p(0 : 1) + \lambda p(0; 0) + \sum_{k=1}^{b} \mu_k p(1; k)$$

$$0 = -(\lambda + \mu_m) p(0; m) + \sum_{k=1}^{b} \mu_k p(m; k), \quad 2 \leq m \leq b$$

$$0 = -(\lambda + \mu_k) p(n; k) + \lambda p(n-1; k), \quad n \geq 1, \quad 1 \leq k \leq b-1$$

$$0 = -(\lambda + \mu_b) p(n; b) + \lambda p(n-1; b) + \sum_{k=1}^{b} \mu_k p(n+b; k), \quad n \geq 1$$

$$0 = p(n; 0), \quad n \geq 1.$$

Indicate how the probabilities $p(0; k), 0 \leq k \leq b-1$ and the marginal probabilities $p(n) = \sum_{k=0}^{b} p(n; k)$ can be obtained from the previous equations. Examine the special case $b = 2, 1/\mu_2 = c/\mu_1, c = 1, 2$. Cosmetatos examines the effect of such a server-sharing rule on performance measures such as the average server utilization, average queue size, and average number of customers in the system and the benefits of the server-sharing scheme (Cosmetatos, 1983c).

4.11. Batch size in systems with general service rule
For an $M/M(a,b)/1$ system, let $\omega = \lambda/(\lambda+\mu)$, $p = \lambda/\mu$. The distribution of service batch size Y is given by

$$g_a = Pr\{Y = a\}$$
$$= 1 - \frac{pP_{0,0}}{\omega(1-r)^2}[-(r-\omega) + (1-\omega)r^{a+1}(a+1-ar)],$$
$$g_y = Pr\{Y = y\}$$
$$= \left(\frac{(1-\omega)}{\omega}\right) yr^y pP_{0,0} \quad a+1 \le y \le b-1,$$
$$g_b = Pr\{Y = b\}$$
$$= \frac{b(r-\omega)}{\omega r(1-r)} pP_{0,0} \quad \text{and}$$
$$E\{Y\} = a\left\{1 + \frac{(r-\omega)pP_{0,0}}{\omega r(1-r)^2} + \frac{(1-\omega)r^{a+1}pP_{0,0}}{(1-r)^3}[1+r+a-ar]\right.$$
$$\left. - \frac{(r-\omega)pP_{0,0}}{\omega(1-r)^3}[(1+2b)+(1-2b)r]\right\}.$$

Replacing ω by $\omega_c = \lambda/(\lambda + c\mu)$, and p by $p_c = \lambda/c\mu$, one gets the corresponding expressions for the $M/M(a,b)/c$ model (Sim and Templeton, 1985).

Note: This gives the distribution of the size of the batch in which a randomly chosen customer is served. This is to be distinguished from the distribution of the size of a randomly chosen batch.

4.12. The model $M/M(1,b)/2$
Show that

$$p^*_{1,0}(s) = \left[\frac{(s+\lambda)}{\mu}\right] p^*_{0,0}(s)$$
$$p^*_{2,0}(s) = \left[\frac{\{(s+\lambda)^2 + s\mu\}}{2\mu^2}\right] p^*_{0,0}(s), \quad n \ge 0.$$

Find $p^*_{0,0}(s)$. Suppose that ($\rho = \lambda/2b\mu < 1$) and that the system is in steady state and r denotes the unique real root in $(0,1)$ of Eq. (4.5.13). Show that the steady-state probabilities are as given in (4.5.15a) to (4.5.17).

4.13. The model $M/M(a,b)/2$.
The busy period may be defined as the interval during which (i) both the servers remain busy or (ii) at least one server remains busy.

Case (i): Show that the busy-period distribution can be obtained from that of the one-channel case by replacing μ by 2μ wherever μ occurs in the one-channel result. Find $b^*(s)$, $f_j^*(0), 0 \leq j \leq a-1$ and the mean and the variance of the busy period.

Case (ii): Here $p_{1,0}(0) = 1$; the busy period is the interval between commencement of service in one of the channels with none in the other channel or queue and the first subsequent epoch when both channels become free. Write down the equations involving

$$p_{ij}^*(s), \quad i = 2, \quad j \geq 0; \quad i = 1, \quad j = 0, 1, \ldots, a-1,$$

and solve them. Denote R as the real root in $(0, 1)$ of Eq. (4.5.13). Denote

$$\begin{aligned} C(s) &= \frac{2\mu p_{2,0}^*(s)}{(s+\lambda+\mu)R - \lambda} \\ &= \frac{[2\mu\lambda^a(1-R)]}{\{(s+\lambda+\mu)^a[(s+\lambda+\mu)R-\lambda][s+2\mu(1-R^a)] \\ &\quad -2\lambda\mu(1-R)[(s+\lambda+\mu)^a R^a - \lambda^a]\}.} \end{aligned}$$

Show that, for $0 \leq q \leq a-1$,

$$p_{1,q}^*(s) = \frac{1}{\lambda}\left(\frac{\lambda}{s+\lambda+\mu}\right)^{q+1} + C(s)\left[R^{q+1} - \left(\frac{\lambda}{s+\lambda+\mu}\right)^{q+1}\right]$$

and hence find the LST of the busy-period distribution. Find the probability that the busy period ends with $q (0 \leq q \leq a-1)$ in the queue, and verify that the total probability equals 1.

4.14. The model $M/M(1,b)/2$.

Define the busy period as the interval during which at least one of the servers remains busy with $p_{1,0}(0) = 1$. Show that the LST $b^*(s)$ of the busy period is given by

$$b^*(s) = \frac{\mu[(s+2\mu) - 2\mu R]}{s(s+\lambda+\mu) + 2\mu(s+\mu)(1-R)}$$

and that the average busy period equals

$$\frac{(1-r) + \left(\frac{\lambda}{2\mu}\right)}{\mu(1-r)}.$$

(R and r refer to the corresponding quantities of the two-channel model.)

Find the corresponding results for an $M/M/2$ queue as a particular case.

4.15. The system $M/M(a,b)/1$.

(a) Let Y be the number of customers that arrive during the period the server is idle with q $(0 \leq q \leq a-1)$ in the queue, and let Z be the number of customers present in the queue at the epoch when the server's idle period commences. Let $P_Y(s)$ and $P_Z(s)$ be the PGF of Y and Z, respectively. Show that

$$P_Y(s) = \frac{s(1-r)(s^a - r^a)}{(1-r^a)(s-r)} \quad \text{and}$$

$$P_Z(s) = \frac{(1-r)\{1-(rs)^a\}}{(1-r^a)(1-rs)}.$$

Find $E(Y)$ and $E(Z)$ and verify that $E(Y) + E(Z) = a$.

(b) Show that the LST of the server's idle period I is given by

$$I^*(s) = \frac{\lambda(1-r)}{(1-r^a)} \frac{1}{(s+\lambda)^a} \frac{\lambda^a - \{r(s+\lambda)\}^a}{\lambda - r(s+\lambda)}.$$

Show further that

$$E(I) = \frac{a(1-r) - r + r^{a+1}}{\lambda(1-r)(1-r^a)} = \frac{1}{\lambda}\left[\frac{a}{1-r^a} - \frac{r}{1-r}\right].$$

Verify that when $a = b = 1, r = \rho$ and $I^*(s) = \frac{\lambda}{(\lambda+s)}$.

(c) Show that

$$p_0 = Pr\{\text{the server is idle}\} = \left[\frac{a(1-r) - r(1-r^a)}{(1-r)^2}\right] p_{0,0}$$

$$= \frac{E(I)}{E(I) + E(B)}.$$

Using the relation, find the fraction of time the server is idle. Show that the average number of batches served during a server busy period is

$$\frac{1}{(1-r^a)}.$$

Deduce the corresponding results for an $M/M/1$ queue.

References and Further Reading

Abramowitz, M., and Stegun, I. A. (Eds.) (1968). *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C.

Bailey, N. T. J. (1954). On queueing processes with bulk service. *J. Roy. Stat. Soc.* **B16**, 80–97.

Borthakur, A. (1971). A Poisson queue with a general bulk service. *J. Ass. Sc. Soc.* **xiv**, 162–167.

Burke, P. J. (1975). Delays in single server queues with batch input. *Opns. Res.* **23**, 830–833.

Chaudhry, M. L., and Templeton, J. G. C. (1983). *A First Course in Bulk Queues*, Wiley, New York.

Chaudhry, M. L., Medhi, J., Sim, S. H., and Templeton, J. G. C. (1987). On a two heterogeneous-server Markovian queue with general bulk service rule. *Sankhyā*, **49** Series B, 36–50.

Cosmetatos, G. P. (1983a). Closed form equilibrium results for the $M/M(\alpha, \infty)/N$ queue. *Euro. J. Opns. Res.* **12**, 203–204.

Cosmetatos, G. P. (1983b). A steady state approximation for the $E_R^X/M/1$ queue. *J. Opnl. Res. Soc.* **34**, 899–902.

Cosmetatos, G. P. (1983c). Increasing productivity in exponential queues by server-sharing. *Omega* **11**, 187–193.

Cromie, M. V., and Chaudhry, M. L. (1976). Analytically explicit results for the queueing system $M^X/M/C$ with charts and tables for certain measures of efficiency. *Oper. Res. Qrtly.* **27**, 733–745.

Easton, G. D., and Choudhry, M. L. (1982). The queueing system $E_k/M(a, b)/1$ and its numerical analysis. *Comp. & Opns. Res.* **9**, 197–205.

Gaver, D. P. (1959). Imbedded Markov chain analysis of a waiting time process in continuous time. *Ann. Math. Stat.* **30**, 698–720.

Ghare, P. M. (1968). Multichannel queueing system with bulk service. *Opns. Res.* **16**, 189–192.

Jackson, R. R. P., and Henderson, J. C. (1966). The time dependent solution to the many server Poisson queue. *Opns. Res.* **14**, 720–722.

Jensen, G. L., Paulson, A. S., and Sullo, P. (1977). Steady state solution for a particular $M^{(k)}/M/1$ queueing system. *Nav. Res. Log. Qrly.* **24**, 651–659.

Kabak, I. W. (1968). Blocking and delay in $M^{(n)}/M/c$ bulk queueing systems. *Opns. Res.* **16**, 830–839.

Luchak, G. (1958). The continuous time solution of the equations of the single channel queue with a general class of service-time distribution by the method of generating functions. *J. Roy. Stat. Soc.* **B 20**, 176–181.

Medhi, J. (1975). Waiting time distribution in a Poisson queue with general bulk service rule. *Management Sci.* **21**, 777–782.

Medhi, J. (1979). Further results in a Poisson queue under a general bulk service rule. *Cahiers Centre d'Etudes de Rech Oper.* **21**, 183–189.

Medhi, J. (1984). *Recent Developments in Bulk Queueing Models*, Wiley Eastern, New Delhi.

Medhi, J. (1994). *Stochastic Processes*, 2nd. ed. Wiley, New York & Wiley Eastern (now New Age Publishers (P) Ltd.), New Delhi.

Medhi, J., and Borthakur, A. (1972). On a two server Markovian queue with a general bulk service. *Cahiers Centre d'Etudes de Rech. Oper.* **14**, 151–158.

Neuts, M. F. (1967). A general class of bulk queues with Poisson input. *Ann. Math. Stat.* **38**, 759–770.

Neuts, M. F. (1979). Queues solvable without Rouchés theorem. *Opns. Res.* **27**, 767–781.

Neuts, M. F. (1981). *Matrix-Geometric Solutions to Stochastic Models—An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore.

Neuts, M. F., and Nadarajan, R. (1982). A multiserver queue with thresholds for the acceptance of customers into service. *Opns. Res.* **30**, 948–960.

Restrepo, R. A. (1965). A queue with simultaneous arrivals and Erlang service distributions. *Opns. Res.* **13**, 375–381.

Reynolds, J. F. (1968). Some results for the bulk arrival infinite server Poisson queue. *Opns. Res.* **16**, 186–189.

Selim, Shokhri, Z. (1997). Time dependent solution and optimal control of a bulk service queue. *J. Appl. Prob.*, **34**, 258–266.

Sim, S. H., and Templeton, J. G. C. (1983). Computational procedures for steady-state characteristics of unscheduled multi-carrier shuttle systems. *Euro. J. Opnl. Res.* **12**, 190–202.

Sim, S. H., and Templeton, J. G. C. (1985). Steady state results for the $M/M(a, b)/c$ batch service system. *Euro. J. Opnl. Res.* **21**, 260–267.

Stadje, W. (1989). Some exact expressions for the bulk arrival queue $M^X/M/1$. *Queueing Systems*, **4**, 85–92.

This Page Intentionally Left Blank

Network of Queues

5.1 Network of Markovian Queues

The queueing models that we have been examining so far are such that every customer or unit demands *one* service and leaves the system as soon as it is obtained. Very often we come across situations where a customer may need more than one service (or different kinds of service) from different servers and may be required to wait before different service channels for service. We can consider two broad types of models. For example, customers at a store may require services in a number of *successive* stages. They may be served at a counter and *then* go to the checkout for payment; a bank customer may first have to go to a manager with his check and then go to the cashier to receive payment. Here each customer has to receive service from two different servers *one after the other* and may have to queue up for service before each of the servers. This is how this kind of model differs from the model with service in several stages, the Erlang-k service. To model such types of queueing situations, one is led to consider queues *in series* or queues *in tandem*. There may be situations in which customers may not be required to go from each service counter to the next one. For example, in maintenance and repair facilities with a number of counters, a particular job may not be required to receive service and thus may not have to pass through each service channel. An outpatient at a hospital may require some, though not every one, of the service facilities, but on the other hand may be required to go to one service again after passing through some other service counter. To model systems of the preceding types, one is required to consider instead of one service channel a *network* of service channels with a separate queue before each of the channels. Modeling of

complex systems often involves such a network of queues. Queueing network models have applications in diverse areas, such as production and assembly lines, maintenance and repair operations, airport terminals, communication networks, computer-sharing and multiprogramming systems, and health-care centers. Before considering general network models, we examine systems in which services are required to be provided in a number of *successive* stages.

5.2 Channels in Series or Tandem Queues

Jackson (1954) was the first to consider queues in series as a model of a queueing system for the overhaul of aircraft engines, where stages of overhaul involve successive operations such as stripping, inspecting, repairing, assembling, and testing. Every unit or customer is served at each of the stages (called phases) one after another and may have to queue up before each service facility, which may have one or more similar parallel servers. The input to each stage (or phase), after the first, is the output from the preceding stage. It is assumed that the queueing space before each service facility is large enough to accommodate any number of waiting customers.

For the sake of simplicity we first consider a system with two phases.

Assume that units arrive in accordance with a Poisson process with rate λ; these units constitute the input to the first stage. Units receive service at the first counter, the service-time distribution being an independent exponential with mean $1/\mu_1$.

The units that emerge from the first stage of service constitute the input to the second stage of service. A unit joins a queue (if needed) before the second service channel, receives service in turn, and then leaves the system. It is assumed that the service-time distribution at the second stage is also independent exponential, with mean $1/\mu_2$. The state of the system at an instant t is given by (n_1, n_2) where n_1 is the number of units in the first phase or before the first service channel and n_2 that before the second channel; let $p(n_1, n_2, t)$ denote the corresponding probability. The system can denoted by the notation $M/M/1 \rightarrow ./M/1$.

The balance equations can be easily written down as follows:

$$p'(n_1, n_2, t) = \lambda p(n_1 - 1, n_2, t) + \mu_1 p(n_1 + 1, n_2, t) + \mu_2 p(n_1, n_2 + 1, t) - (\lambda + \mu_1 + \mu_2) p(n_1, n_2, t), \quad n_1 \geq 1, \quad n_2 \geq 1, \tag{5.2.1}$$

$$p'(n_1, 0, t) = \lambda p(n_1 - 1, 0, t) + \mu_2 p(n_1, 1, t) - (\lambda + \mu_1) p(n_1, 0, t), \quad n_1 \geq 1, \tag{5.2.2}$$

$$p'(0, n_2, t) = \mu_1 p(1, n_2 - 1, t) + \mu_2 p(0, n_2 + 1, t) - (\lambda + \mu_2) p(0, n_2, t), \quad n_2 \geq 1, \tag{5.2.3}$$

$$p'(0, 0, t) = -\lambda p(0, 0, t) + \mu_2 p(0, 1, t). \tag{5.2.4}$$

Assume that $\rho_1 = \lambda/\mu_1 < 1, \rho_2 = \lambda/\mu_2 < 1$. Then the steady state is reached for large t, and $\lim_{t\to\infty} p(n_1, n_2, t)$ exists. Denote $\lim_{t\to\infty} p(n_1, n_2, t) = p(n_1, n_2)$. Then it can be shown that

$$p(n_1, n_2) = \rho_1^{n_1}\rho_2^{n_2}\, p(0,0) \quad \text{and}$$
$$p(0,0) = (1-\rho_1)(1-\rho_2).$$

Thus,

$$p(n_1, n_2) = \left[(1-\rho_1)\rho_1^{n_1}\right]\left[(1-\rho_2)\rho_2^{n_2}\right]. \tag{5.2.5}$$

The relation (5.2.5) indicates that in steady state each phase behaves *independently* of the other. Further, the second phase behaves like a system with an input process that is Poisson with rate λ. This also follows from Burke's theorem on the output process of an $M/M/1$ queue in steady state. (See Section 3.2.3.) The input process to the second phase is an *independent* Poisson process with rate λ. The second phase behaves as an $M/M/1$ queue independent of the behavior of the first phase. Hence, it follows that in steady state (when $\rho_1 < 1, \rho_2 < 1$), $p(n_1, n_2)$ is given by (5.2.5). The result can be extended to a finite number of simple $M/M/1$ queues in series or in tandem.

Example 5.1. Suppose that k number of $M/M/1$ queues in tandem are in steady state, with λ as the arrival rate and μ_i as the service rate of the ith phase. Find

(1) the probability $P(n_1, \ldots, n_k)$ that there are n_1 in the first phase, n_2 in the second phase, and so on;

(2) the expected number in the ith phase; and

(3) the expected number in the complete system.

For $\mu_1 = \mu_2 = \cdots = \mu_k$ find the probability that there are n units in the complete system.

Since each of the phases behaves independently of the others

$$P(n_1, \ldots, n_k) = \prod_{i=1}^{k}(1-\rho_i)\rho_i^{n_i}, \quad \text{where } \rho_i = \frac{\lambda}{\mu_i}.$$

The ith phase is an $M/M/1$ queue with rates λ and μ_i and so the expected number in the system in the ith phase is given by

$$\frac{\rho_i}{1-\rho_i}$$

and the expected number in the complete system is given by

$$\sum_{i=1}^{k} \frac{\rho_i}{1-\rho_i}.$$

Suppose that $\mu_1 = \cdots = \mu_k$—that is, $\rho_1 = \cdots = \rho_k = \rho$. The number of distinguishable arrangements in which the total number of n (indistinguishable) things can be put in k cells is

$$\binom{n+k-1}{k-1}.$$

Hence, the probability that there are (a total of) n units in the complete system is given by

$$\binom{n+k-1}{k-1}(1-\rho)^k\rho^n. \tag{5.2.6}$$

5.2.1 Queues in series with multiple channels at each phase

Suppose that there are k service channels in series in steady state. The arrivals to the first phase after completing service there proceed to the second and so on and finally emerge from the system after having service at the kth channel. Suppose that phase i behaves as an $M/M/c_i$ queue and that there are *ample* holding spaces in front of each queue. In steady state the output process of each phase is Poisson with rate equal to the initial input rate λ. From Burke's theorem we find that phase-by-phase decomposition is possible in such a case. The probability $P(n_1, \ldots, n_k)$ that there are n_i units in the system in the ith phase (which is an $M/M/c_i$ queue, $i = 1, 2, \ldots, k$) is therefore given by

$$\begin{aligned} P(n_1, \ldots, n_k) &= P_1(n_1) \cdots P_k(n_k) \\ &= \prod_{i}^{k} P_i(n_i), \end{aligned}$$

where $P_i(n_i) = Pr\{\text{there are } n_i \text{ in the system in an } M/M/c_i \text{ queue in steady state}\}$.

The result is obtained in a *product form*. We shall see that such product-form results also hold good even in more general cases of network of queues.

So far studies have been confined to networks of *Markovian* queues only, such that each service center is an $M/M/c_i$ queueing system. Further, it is assumed that *steady state* exists, a sufficient condition for which is that each traffic intensity $\rho_i = \lambda/\mu_i < 1$. In the usual terminology of *networks of queues*, each phase or service center is called a *node*. A node may be one with one or more than one identical servers. In the more general system, customers enter

the system at various nodal points, join the queue (if any), and receive service. In a tandem queue, the customers proceed sequentially from one node to the next one, and in a feedforward queue customers proceed from one node to the next node in the forward direction.

Again we may have an *open* network where a customer after completing his service at a node may leave the system. On the other hand, in a closed network there is a fixed and finite number of customers—say, K—such that they can neither leave the system nor be joined from outside by others. *A closed network in series or tandem* in which the customers after completing service at the last node again go back and join the first node is called a *cyclic* network.

Jackson (1957) considered a more general network of queues, which is known as a *Jackson network*. There are various situations that can be modeled by some of these network patterns. The idea of modeling a job shop as a queueing network is old. More recently, similar ideas have appeared in models of computer systems, road-traffic systems, command and control systems, data-transmitting systems, teletraffic systems, and others. The applications of queueing networks also extend to biology (neural networks), disease (compartmental models), and polymerization (cluster models). Kelly (1979) gives many varied examples of queueing networks. All these applications make the study of queueing networks more significant.

Remark 1: We have assumed so far that the waiting space before each of the servers in series is large enough. However, this assumption may not always be valid. Use the notation $M/M/1/L_1 + 1$ to indicate that the server has before him a waiting space for L_1 customers besides the one being served, if any (i.e., in all $L_1 + 1$ places). We may have the following three types of disciplines.
(a) The customer at the server j who, upon completion of service, cannot gain immediate access to the server $(j + 1)$ overflows the server $(j + 1)$ and goes immediately to server $j + 2$. This is called an *overflow discipline.*
(b) A customer who cannot gain immediate access to the server $(j + 1)$ stays idle at server j until a space before the server $(j + 1)$ becomes available. This is called a *blocking discipline.*

Consider the system $M/M_1/1 \to ./M_2/1/L_2 + 1 (L_2 < \infty)$ with loss discipline (finite waiting space before the second server). The service-time distribution of server i is exponential, with mean $1/\mu_i$ and $\rho_i = \lambda/\mu_i$, λ being the arrival rate to server 1. Here the queue-length processes are independent and

$$P(n_1, n_2) = \left[(1 - \rho_1)\rho_1^{n_1}\right] \cdot \left[\frac{(1 - \rho_2)\rho_2^{n_2}}{(1 - \rho_2)^{L_2+2}}\right],$$
$$n_1 = 0, 1, \ldots, \quad n_2 = 0, 1, \ldots, L_2 + 1.$$

The system has a product-form solution. The system

$$M/M_1/1/L_1 \to ./M_2/1/L_2$$

with $L_i < \infty$ has a blocking discipline that imposes dependence between queue-length processes before server 1 and server 2. The system does not have a product-form solution. There are a large number of variations for the series network of queues, and many papers have been devoted to them.

Remark 2: Chen (1989) considers two single-server tandem queues, the first being an $M/D/1$ queue and the second having exponential service times—that is, the tandem queue $M/D/1 \rightarrow ./M/1$. He shows that the steady-state distribution $p(n_1, n_2)$ does not have a product form and hence that the two queues are not independent, and that the sojourn times of the two queues are not independent.

Remark 3: If the input is Poisson, and service times at all the channels are exponential, then the output is Poisson. Further sojourn times at different channels are independent irrespective of the order of the channels. Weber (1979) shows that the final departure process is independent of the order of the channels for an arbitrary arrival process as long as the servers are exponential.

Remark 4: Boxma (1986) gives a review of results relating to (1) two queues in series and (2) two parallel queues with a single server. For references, see Boxma (1986) and Gnedenko and König (1983).

5.3 Jackson Network

In this model, customers from one node i proceed to an arbitrary node, and fresh customers may join a node from outside. Suppose that there are k nodes, where the ith node ($i = 1, \ldots, k$) consists of c_i exponential servers with parameter μ_i. Customers after receiving service at the ith node proceed to the jth node with probability p_{ij}. Suppose further that the ith node may receive customers from a Poisson stream with rate λ_i from outside the system as well. Customers at node i depart from the system with probability $q_i = 1 - \sum_{j=1}^{k} p_{ij}$.

This is Jackson's network model. The model is very general. For example, with $\lambda_i = 0$, and $p_{ij} = 1$, wherever $j = i + 1, 1 \leq i \leq k - 1$, and $q_k = 1$, we get k queues in tandem. For a closed network with k members. $\lambda_i = 0, \sum_{j=1}^{k} p_{ij} = 1$ so that $q_i = 0$ for each i. For a cyclic network, $\lambda_i = 0$, $p_{ij} = 1$, $j = i + 1, 1 \leq i \leq k - 1$ and $p_{k1} = 1$ (and $q_i = 0$ for all i).

Consider Jackson's general network model with k nodes. The input to the ith node consists of outputs of the other nodes as well as the external input λ_i. The total average arrival rate of customers α_i to node i is the sum of the Poisson arrival rates λ_i from outside the system plus the arrival rate from arrivals to node i from (other) internal nodes $\sum_j p_{ji}\alpha_j$.

Figure 5.1 provides a diagram of node i.

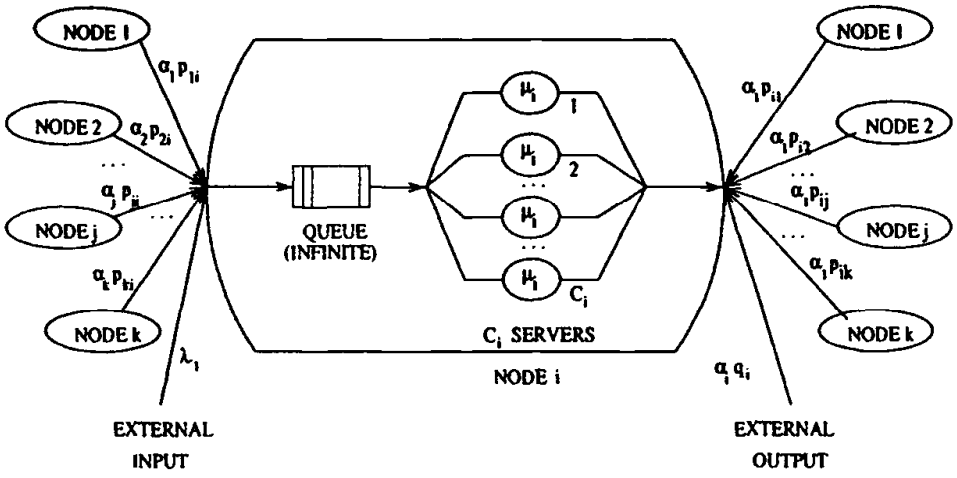


Figure 5.1 Node i in a Jackson Network.

It easily can be seen that the parameters α_i satisfy the equation

$$\alpha_i = \lambda_i + \sum_{j=1}^{k} p_{ji}\alpha_j, \quad i = 1, 2, \ldots, k; \tag{5.3.1}$$

α_i gives the effective arrival rate to the node i or effective rate of flow through node i.

The preceding equations are known as *traffic equations, flow balance equations, or conservation equations* for the rate of flow through an arbitrary node. Note that $\alpha_i q_i$ is the rate of departure from the system from the node.

The existence of the solution of the preceding set of equations is a necessary condition for the existence of the steady-state distribution in a Jackson network.

In his fundamental paper, Jackson (1957) shows that for a Jackson network of (Markovian) queues, the particular product-form result of marginal distribution holds in equilibrium, implying the independence of various nodes in the network. This is commonly referred to as Jackson's theorem, which we consider next. The proof given by Jackson is straightforward; it uses the usual probability arguments employed to derive balance equations. It is assumed that the network is completely open.

Theorem 5.1. (Jackson's Theorem)
Let $(n_1, n_2, \ldots, n_k)$ denote the state of the complete system in which there are n_i (in the queue and in service) at node i in a Jackson network of Markovian queues in equilibrium and let $p(n_1, \ldots, n_k)$ be the probability that the system is in the state $(n_1, \ldots, n_k)$.

Assume that

$$\rho_i = \frac{\alpha_i}{\mu_i} < 1, \quad i = 1, 2, \ldots, k,$$

where $\{\alpha_i\}$ *are given by the balance equations*

$$\alpha_i = \lambda_i + \sum_j \alpha_j p_{ji}, \quad i = 1, 2, \ldots, k.$$

If $p_i(n_i)$ *denotes the probability that there are* n_i *in the system (in queue plus service) for the* $M/M/c_i$ *queue with input rate* α_i, *and service rate* μ_i *for each of the* c_i *servers,*

$$\begin{aligned} p_i(n) &= p_i(0)\frac{\left(\frac{\alpha_i}{\mu_i}\right)^n}{n!}, \quad n = 0, 1, 2, \ldots, c_i \\ &= p_i(0)\frac{\left(\frac{\alpha_i}{\mu_i}\right)^n}{\left[c_i! c_i^{n-c_i}\right]}, \quad n = c_i + 1, \ldots \end{aligned} \tag{5.3.2}$$

Then

$$p(n_1, \ldots, n_k) = p_1(n_1)\, p_2 \ldots p_k(n_k). \tag{5.3.3}$$

Proof: Let $p_t(n_1, \ldots, n_k)$ be the probability that the complete system is in state $(n_1 \ldots, n_k)$ at time t; then considering in the usual manner the infinitesimal interval $(t, t + h)$ following the interval $(0, t)$, we can write the differential equation satisfied by p_t. Denote

$$\begin{aligned} q_i = 1 - \sum_j p_{ij}, \quad a_i(n) = \min\{n, c_i\} &= n, \quad \text{if } n < c_i \\ &= c_i \quad \text{if } n \geq c_i \\ \delta_i = \min\{n_i, 1\} &= 1, \quad n_i \geq 1 \\ &= 0, \quad n_i = 0. \end{aligned}$$

The state at $t + h$ can be reached from $(n_1, \ldots, n_k)$, the state at t, in one of the following four mutually exclusive ways.

(A) State at t is $(n_1, \ldots, n_k)$, and no arrival occurs to any node from, external source, nor does any departure occur from any node in the interval $(t, t + h)$ of length h. We get

$$Pr(A) = p_t(n_1, \ldots, n_k)\left[1 - \left(\sum \lambda_i\right)h - \sum a_i(n_i)\mu_i h\right] + o(h). \tag{5.3.4}$$

(B) State at t is $(n_1, \ldots, n_i + 1, \ldots, n_k)$ for some $i (i = 1, \ldots, k)$, there is one service completion at that node i in $(t, t + h)$, and this completion departs from the system (with probability q_i). We get

$$Pr(B) = \sum_{i=1}^{k} p_t(n_1, \ldots, n_i + 1, \ldots, n_k)[a_i(n_i + 1)h\mu_i]q_i + o(h). \tag{5.3.5}$$

(C) State at t is $(n_1, \ldots, n_i - 1, \ldots, n_k)$ for some $i (i = 1, 2, \ldots, k)$ and there is one arrival from the external source to node i in the interval $(t, t + h)$. We get

$$Pr(C) = \sum_{i=1}^{k} p_t(n_1, \ldots, n_i - 1, \ldots, n_k)\{\lambda_i h \delta_i\} + o(h). \tag{5.3.6}$$

(Here $\delta_i = 1$, if $n_i \geq 1$, and $\delta_i = 0$ if $n_i = 0$.)
(D) State at t is $(n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_k)$, there is one service completion at node i in $(t, t + h)$, and the one whose service is completed moves to node j with probability p_{ij}. Thus,

$$\begin{aligned} Pr(D) = \sum_i \sum_j & p_t(n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_k) \\ & \times [a_i(n_i + 1)\mu_i h p_{ij}] + o(h). \end{aligned} \tag{5.3.7}$$

Thus,

$$p_{t+h}(n_1, \ldots, n_k) = Pr(A) + Pr(B) + Pr(C) + Pr(D). \tag{5.3.8}$$

Following the usual procedure of transferring the term $Pr(A)$ to the left-hand side, then dividing both sides by h, and finally taking limits as $h \to 0$, we get the differential-difference equation for $p'(t)$.

Putting $p_t(n_1, \ldots, n_k) = p(n_1, \ldots, n_k)$ and $p'(t) = 0$, we get the equations satisfied by the steady-state probabilities as follows:

$$\begin{aligned} & \left[\sum_i \lambda_i + \sum_i a_i(n_i)\mu_i \right] p(n_1, \ldots, n_k) \\ & \quad = \sum_i a_i(n_i + 1)\mu_i q_i \, p(n_1, \ldots, n_i + 1, \ldots, n_k) \\ & \qquad + \sum_i \lambda_i \delta_i \, p(n_1, \ldots, n_i - 1, \ldots, n_k) \\ & \qquad + \sum_i \sum_j a_i(n_i + 1)\mu_i p_{ij} \, p(n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_k). \end{aligned} \tag{5.3.9}$$

We shall now show that $p(n_1, \ldots, n_k)$ given by (5.3.3) is a solution of the preceding equation. From the form of $p(n_1, \ldots, n_k) = p_1(n_1) \ldots p_k(n_k)$, (given in (5.3.3)) and the form of $p_i(n_i)$ (given in (5.3.2)), we get

$$\begin{aligned} \frac{p(n_1, \ldots, n_i + 1, \ldots, n_k)}{p(n_1, \ldots, n_i, \ldots, n_k)} &= \frac{p_i(n_i + 1)}{p_i(n_i)} \\ &= \frac{\alpha_i}{\mu_i a_i(n_i + 1)} \end{aligned} \tag{5.3.10}$$

$$\frac{p(n_1,\ldots,n_i-1,\ldots,n_k)}{p(n_1,\ldots,n_i,\ldots,n_k)} = \frac{\mu_i a_i(n_i)}{\alpha_i} \tag{5.3.11}$$

and

$$\frac{p(n_1,\ldots,n_i+1,\ldots,n_j-1,\ldots,n_k)}{p(n_1\ldots,n_i,\ldots,n_j,\ldots,n_k)} = \frac{\alpha_i}{\mu_i a_i(n_i+1)}\,\frac{\mu_j a_j(n_j)}{\alpha_j}. \tag{5.3.12}$$

Dividing both sides of (5.3.9) by $p(n_1,\ldots,n_i,\ldots,n_k)$ and using (5.3.10), (5.3.11), and (5.3.12), we get

$$\begin{aligned}\left[\sum_i \lambda_i + \sum_i a_i(n_i)\mu_i\right] &= \sum_i a_i(n_i+1)\mu_i q_i \left[\frac{\alpha_i}{\mu_i a_i(n_i+1)}\right] \\ &\quad + \sum_i \frac{\lambda_i \delta_i \mu_i a_i(n_i)}{\alpha_i} \\ &\quad + \sum_i \sum_j a_i(n_i+1)\mu_i p_{ij}\,\frac{\alpha_i \mu_j a_j(n_j)}{\alpha_j \mu_i a_i(n_i+1)}.\end{aligned} \tag{5.3.13}$$

We have

$$\begin{aligned}\sum_i \alpha_i q_i &= \sum_i \alpha_i \left[1 - \sum_{j=1}^{k} p_{ij}\right] \\ &= \sum \alpha_i - \sum_j \left(\sum_i \alpha_i p_{ij}\right) \\ &= \sum_i \alpha_i - \sum_j (\alpha_j - \lambda_j) = \sum_j \lambda_j\end{aligned} \tag{5.3.14}$$

$$\begin{aligned}\sum_i \sum_j \frac{\alpha_i \mu_j a_j(n_j) p_{ij}}{\alpha_j} &= \sum_j \frac{a_j(n_j)}{\alpha_j}\,\mu_j \left\{\sum_i \alpha_i p_{ij}\right\} \\ &= \sum_j \frac{a_j(n_j)}{\alpha_j}\,\{\mu_j(\alpha_j - \lambda_j)\} \\ &= \sum_j \mu_j a_j(n_j) - \sum_j \frac{\lambda_j \mu_j a_j(n_j)}{\alpha_j} \\ &= \sum_i \mu_i a_i(n_i) - \sum_i \frac{\lambda_i \mu_i a_i(n_i)}{\alpha_i}.\end{aligned} \tag{5.3.15}$$

Noting that $\delta_i a_i(n_i) = a_i(n_i)$, we get that the RHS of (5.3.13) is equal to

$$\sum \lambda_i + \sum \frac{\lambda_i \mu_i a_i(n_i)}{\alpha_i} + \sum \mu_i a_i(n_i) - \sum \frac{\lambda_i \mu_i a_i(n_i)}{\alpha_i}$$

$$= \sum \lambda_i + \sum \mu_i a_i(n_i)$$

$$= \text{LHS of Eq. (5.3.13).}$$

This shows that (5.3.3) (with (5.3.2)) satisfies (5.3.9)—in other words, (5.3.3) is a solution of (5.3.9). It can be verified by direct summation that (5.3.3) also satisfies the normalizing condition $\sum_{n_i} p(n_1, \ldots, n_k) = 1$, $(n_i \geq 0)$.

The justification that (5.3.3) gives the steady-state distribution follows from a result of Markov process theory that can be stated as follows.

If there is a positive solution of the balance equations of an irreducible Markov process and if such a solution satisfies the normalizing condition, then the steady-state distribution of the process exists and is given by that solution. This establishes the uniqueness of the solution. This completes the proof. ■

Note 1: An alternative proof of Jackson's theorem (using the concept of partial balance) has been given by Lemoine (1977).

Note 2: The equilibrium-queue-length distribution in a Jackson network is of the product form.

Note 3: Jackson's theorem contains the remarkable result that whenever the equilibrium condition exists, each node in the network behaves as if it were an independent $M/M/c_i$ queue with Poisson input. It is implied that the RV n_i representing the number of states in individual nodes i $(i = 1, 2, \ldots, k)$ in steady state are independent random variables.

Note 4: In equilibrium, the external departure streams or outputs from individual nodes are independent Poisson processes with rate $q_i \alpha_i$ for node $i, i = 1, 2, \ldots, k$. The relation (5.3.14) shows that $\sum \lambda_i = \sum \alpha_i q_i$. That is, under equilibrium the total *external* input flow rate is equal to the total external output flow rate. However, the arrival and departure processes at an $M/M/1$ node are not always equivalent (Walrand, 1982a).

Note 5: In general, the total *input* process into an individual node is not necessarily Poisson, even under equilibrium conditions. For a demonstration of this, see Gelenbe and Mitrani (1980, pp. 85–86). Here non-Poisson total arrivals or inputs see time averages, and this accounts for the simple state probability results.

Note 6: The traffic equations

$$\alpha_i = \lambda_i + \sum_j \alpha_j p_{ji}, \quad i = 1, \ldots, k$$

have a unique solution. Writing $\boldsymbol{\alpha} = (\alpha_1 \ldots, \alpha_k), \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)\, \boldsymbol{P} = (p_{ij})$, we get

$$\boldsymbol{\alpha} = \boldsymbol{\lambda} + \boldsymbol{\alpha P} \quad \text{or} \quad \boldsymbol{\alpha}(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{\lambda}. \tag{5.3.16}$$

Now, since in an open network any customer in any queue eventually leaves the network, it follows that each element of $\boldsymbol{P}^n$ converges to 0 as $n \to \infty$. Thus, $(\boldsymbol{I} - \boldsymbol{P})^{-1}$ converges; that is $(\boldsymbol{I} - \boldsymbol{P})$ has an inverse, so that the rank of $(\boldsymbol{I} - \boldsymbol{P})$ is k. This demonstrates that $\boldsymbol{\alpha} = \boldsymbol{\lambda} + \boldsymbol{\alpha P}$, for given $\boldsymbol{\lambda}$ and $\boldsymbol{P}$ has a unique solution in $\boldsymbol{\alpha}$, $\boldsymbol{P}$ is called the *transfer, switching,* or *routing* matrix. The square matrix $\boldsymbol{P}$ from an open network is not stochastic.

If the outside world (the source from which inputs come and the sink to which external outputs go) is considered as node "O," then the square matrix with one more column and row will be stochastic.

Note 7: Consider a state $\boldsymbol{N} = (n_1, \ldots, n_i, \ldots, n_k)$ of the network in equilibrium. For each node i, the rate of flow out of state $\boldsymbol{N}$ due to a departure of a customer from node i is equal to the rate of flow into the state $\boldsymbol{N}$ due to the arrival of a customer into node i due to either external input or internal transfer. This gives *local-balance* relations; Eq. (5.3.9) involving the rates of flow of the complete network constitutes the *global-balance* equation.

The local-balance equation can be obtained by equating the rate of flow out of state $(n_1, \ldots, n_i, \ldots, n_k)$ due to a customer leaving node i with the rate of flow into state $(n_1, \ldots, n_i, \ldots, n_k)$ due to arrival of a customer to node i, either from outside or from some of the other internal nodes. We have

$$\text{rate of flow out of } i = a_i(n_i)\mu_i\, p(n_1, \ldots n_i, \ldots, n_k)(1 - p_{ii}) \quad \text{and} \tag{A}$$

$$\text{rate of flow into of } i = \lambda_i \delta_i\, p(n_1 \ldots, n_i - 1, \ldots, n_k) \quad \text{(from outside)} \tag{B}$$

$$= \sum_{j \neq i} a_j(n_j + 1)\mu_j\, p_{ji}\, p(n_1, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_k). \tag{C}$$

Assuming that p is of the form (5.3.3), and using (5.3.10)–(5.3.12), we get the following form of local-balance equation:

$$a_i(n_i)\mu_i(1 - p_{ii}) = \frac{\lambda_i \delta_i \mu_i a_i(n_i)}{\alpha_i} + \sum_{j \neq i} \frac{\alpha_j\, p_{ji}\, a_i(n_i)\mu_i}{\alpha_i}.$$

Note 8: In notes (3) and (5) we observe that, although the equilibrium distribution has the product form as if the service facilities were independent and all arrival processes were Poisson, the arrival process within the network is, in general, not Poisson. Further, the facilities and the associated queue-length processes (time-dependent) are not independent (Melamed, 1979).

This shows that equilibrium distribution does not capture the transient or time-dependent behavior. Similar behavior is also noted in the case of a

$G/M/1$ queue. This points to the limitations of the steady-state result. (See also Section 2.5.)

Note 9: Goodman and Massey (1984) have generalized Jackson's theorem to the nonergodic case. Their results completely characterize the large-time behavior ($t \to \infty$) of Jackson networks.

5.4 Closed Markovian Network (Gordon and Newell Network)

In contrast to Jackson's open network, Gordon and Newell (1967a) considered a *closed* network of Markovian queues, in which a fixed and finite number of customers—say, K—circulate through the network, there being no external input or departure from the network. This corresponds to Jackson's model in which $\lambda_i = 0$ and $q_i = 0$ for each i and $n_1 + n_2 + \cdots + n_k = K$ (fixed).

With $\lambda_i = 0$, the traffic equation reduces to $\boldsymbol{\alpha} = \boldsymbol{\alpha}\boldsymbol{P}$ and $q_i = 0$ for each i. This implies that the switching matrix $\boldsymbol{P}$ is stochastic. Assuming that $\boldsymbol{P}$ is irreducible, one can therefore regard $\boldsymbol{\alpha}$ as the unique stationary distribution of a discrete parameter Markov chain with transition probability matrix $\boldsymbol{P}$. For this closed network $\boldsymbol{N}(t) = (n_1(t), \ldots, n_k(t))$ is an irreducible continuous-time Markov chain with finite state space and therefore possesses an equilibrium distribution.

Theorem 5.2. (Gordon and Newell, 1967a)
Consider a closed network with k nodes within which a total of K customers (jobs) circulate according to the routing matrix $\boldsymbol{P}$ *(of order k). The node i has* C_i *identical exponential servers with rate* $\mu_i, i = 1, 2, \ldots, k$.

Let $\boldsymbol{\alpha}'$ *be any nonzero solution of* $\boldsymbol{\alpha}' = \boldsymbol{\alpha}'\boldsymbol{P}$ *and*

$$\rho_i = \frac{\alpha_i}{\mu_i}, \quad d_i(n_i) = \begin{cases} n_i!, & n_i \le c_i \\ (c_i!)(c_i)^{n_i - c_i}, & n_i > c_i. \end{cases} \tag{5.4.1}$$

Then the steady-state probability that the state of the system is $(n_i, \ldots, n_k)$ *is given by*

$$p(n, \ldots, n_k) = \frac{1}{A(K)} \prod_{i=1}^{k} \frac{\rho_i^{n_i}}{d_i(n_i)}, \tag{5.4.2}$$

where

$$A(K) = \sum_{\substack{n_i \ge 0 \\ \sum n_i = K}} \prod_{i=1}^{k} \frac{\rho_i^{n_i}}{d_i(n_i)}, \tag{5.4.3}$$

$[(A(K)]^{-1}$ *being the normalizing constant).*

Proof: The proof of the theorem is also by direct verification—that is, by showing that (5.4.2) is a solution of the corresponding balance equation (obtained by putting $\lambda_i = 0$ and $q_i = 0$ in Eq. (5.3.9)). ■

Particular case. For $c_i = 1$ for all i (each node having only one server) $d_i(n_i) = 1$.

Notes:

(1) The equilibrium distribution is still of the product form. The product, however, is not the product of distributions of states of the individual nodes. The states of the nodes are not independent, as is evident also from the fact that $\sum n_i = K$ (fixed).

Compare this with the corresponding result for a Jackson (open) network.

(2) The routing matrix $\boldsymbol{P}$ is a stochastic matrix.

(3) Kiessler (1989) gives a simple proof of the equivalence of input and output intervals (interinput and interoutput distributions) at each node in a Jackson network of single-server nodes. He also shows that the proof is valid for a closed (Gordon-Newell) network of single-server nodes.

Example 5.2. Two-node system with feedback

Consider a closed network having two nodes and K jobs circulating among these two nodes (node 1 corresponding to CPU and node 2 to an I/O device of a computer system having a fixed number of K circulating programs). Suppose that the lengths of successive service times at node i (CPU execution bursts and I/O bursts) are IID exponential with mean $1/\mu_i, i = 1, 2$. At the end of service at node 1, a customer goes to node 2 with probability p or is fed back to node 1 with probability $q = 1 - p$. (At the end of a CPU burst, a program requests an I/O operation with probability p.) Let n be the number of customers at node 1 (programs in the CPU queue) including the one being served, so that $K - n$ is the number at node 2 (in the I/O queue), and let $p(n, K - n)$ be the steady-state probability. Here

$$P = \begin{pmatrix} q & p \\ 1 & 0 \end{pmatrix}$$

and $\boldsymbol{\alpha}' = \boldsymbol{\alpha}'\boldsymbol{P}$ leads to $\boldsymbol{\alpha}' = (a, ap)$ where a is an arbitrary constant. We have $\rho_1 = a/\mu_1$ and $\rho_2 = ap/\mu_2$. We have

$$\begin{aligned} p(n, K - n) &= \frac{1}{A(K)} \left(\frac{a}{\mu_1}\right)^n \left(\frac{ap}{\mu_2}\right)^{K-n} \\ &= \frac{1}{A(K)} \left(\frac{1}{r}\right)^K r^n, \end{aligned}$$

where

$$r = \frac{\mu_2}{p\mu_1} \quad \text{and}$$

$$A(K) = \sum_{n=0}^{K} \left(\frac{1}{r}\right)^K r^n = \left(\frac{1}{r}\right)^K \cdot \frac{1 - r^{K+1}}{1 - r}, \quad r \neq 1$$

$$= \left(\frac{1}{r}\right)^K \cdot \frac{1}{K+1}, \quad r = 1.$$

Thus, for $n = 0, 1, \ldots, K$,

$$p(n, K-n) = \frac{(1-r)r^n}{1 - r^{K+1}}, \quad r \neq 1,$$

$$= \frac{1}{K+1}, \quad r = 1. \tag{5.4.4}$$

In other words,

$$p(K-m, m) = \frac{\left[1 - \left(\frac{1}{r}\right)\right]\left(\frac{1}{r}\right)^m}{1 - \left(\frac{1}{r}\right)^{K+1}}, \quad r \neq 1,$$

$$= \frac{1}{K+1}, \quad r = 1. \tag{5.4.5}$$

The server-utilization factors u_1 and u_2 at node 1 (CPU) and at node 2 (I/O), respectively, are given by

$$u_1 = \sum_n p(n, K-n) = 1 - p(0, K) = \frac{r - r^{K+1}}{1 - r^{K+1}} = \frac{r(1 - r^K)}{1 - r^{K+1}}, \quad r \neq 1$$

$$= \frac{K}{K+1}, \quad r = 1, \quad \text{and} \tag{5.4.6}$$

$$u_2 = \sum_n p(n, K-n) = 1 - p(K, 0) = \frac{\frac{1}{r} - \left(\frac{1}{r}\right)^{K+1}}{1 - \left(\frac{1}{r}\right)^{K+1}} \tag{5.4.7}$$

$$= \frac{1 - r^K}{1 - r^{K+1}}, \quad r \neq 1$$

$$= \frac{K}{K+1}, \quad r = 1. \tag{5.4.8}$$

Notes:

(1) We have

$$(1) \ \frac{u_1}{u_2} = r = \frac{\mu_2}{p\mu_1} = \frac{\rho_1}{\rho_2} \quad \text{and}$$

$$(2) \ u_1 = \frac{A(K-1)}{A(K)}. \tag{5.4.9}$$

(2) Queueing networks serve as models for multiprogrammed computer systems and communication networks and certain parts-manufacturing systems. The number K of circulating programs in a computer network is known as the *level (or degree) of multiprogramming.*

5.5 Cyclic Queue

Consider a closed network of K nodes such that the output of the node i goes to the next node $i + 1 (1 \leq i \leq k - 1)$, whereas the output of the last node k feeds back to node 1 and so on. Such a queue is called a *cyclic queue.* A cyclic queue is a special kind of closed-network queue having routing matrix

$$\boldsymbol{P} = (p_{ij}),$$

where

$$\begin{aligned} p_{ij} &= 1, \quad j = i + 1, \quad 1 \leq i \leq k - 1 \\ &= 1, \quad i = k, \quad j = 1 \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

The results of a closed-queueing network will apply. The solution of $\boldsymbol{\alpha}' = \boldsymbol{\alpha}' \boldsymbol{P}$ leads to

$$\begin{aligned} \rho_1 \mu_1 &= \rho_k \mu_k \\ \rho_i \mu_i &= \rho_{i-1} \mu_{i-1}, \quad i = 1, 2, \ldots, k \\ &\left(\rho_i = \frac{\alpha_i}{\mu_i} \right) \end{aligned}$$

so that

$$\begin{aligned} \rho_i &= \frac{\mu_k}{\mu_i} \rho_k, \quad i = 1 \\ &= \frac{\mu_{i-1}}{\mu_i} \rho_{i-1}, \quad i = 2, 3, \ldots, k. \end{aligned}$$

Thus,

$$\rho_2 = \left(\frac{\mu_1}{\mu_2} \right) \rho_1, \quad \rho_3 = \left(\frac{\mu_2}{\mu_3} \right) \rho_2 = \left(\frac{\mu_1}{\mu_3} \right) \rho_1$$

$$\rho_k = \left(\frac{\mu_1}{\mu_k} \right) \rho_1.$$

Using Theorem 5.2 (with $c_i = 1$), we get

$$p(n_1 \ldots, n_k) = \frac{1}{A(K)} \prod_{i=1}^{k} \rho_i^{n_i}$$

$$= \frac{1}{A(K)} \rho_1^{n_1} \frac{(\mu_1)^{K-n_1}}{(\mu_2)^{n_2} \cdots (\mu_k)^{n_k}}.$$

Thus, we get

$$p(n_1 \ldots, n_k) = \frac{1}{A_1(K)} \frac{\mu_1^{K-n_1}}{(\mu_2)^{n_2} \cdots (\mu_k)^{n_k}}$$

where $[A_1(K)]^{-1}$ is the normalizing constant. $[A_1(K)]^{-1}$ equals the sum of the second factor on the RHS over $n_i s$ such that $\sum_{i=1}^{k} n_i = K$.

Notes:

(1) The corresponding result for multiple servers at the nodes can be written down easily.

(2) Efficient and stable computational algorithms are put forward for calculation of the normalization constant—for example, see Buzen (1973) and Allen (1990).

Example 5.3. Consider a cyclic queue with two nodes and K circulating jobs. The first node (node 1) has an exponential server with rate μ, and the second node (node 2) has an exponential server with rate λ. (This can be treated as a special case of Example 5.2 with $\mu_1 = \mu, \mu_2 = \lambda$, and $p = 1$.) We have

$$r = \frac{\mu_2}{p\mu_1} = \frac{\lambda}{\mu} = \rho \text{ (say)} \quad \text{and}$$

$$p(n, K-n) = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, \quad \rho \neq 1 \tag{5.5.1}$$

$$= \frac{1}{K+1}, \quad \rho = 1.$$

Now for an $M/M/1/K$ queue, the steady-state probability that there are $n(\leq K)$ in the system is also given by (5.5.1). Thus, a two-node cyclic queue as described previously may be considered as equivalent to a limited-space $M/M/1/K$ queue. (See Problem 5.6 for a more general result.) The result was also discussed in Section 3.3.3.

Remarks:

(1) Distributions of sojourn time and cycle time in a cyclic exponential network have been considered, as noted in concluding remarks. See Problems 5.9 and 5.10 for some specific cases where distributions are of product form.

(2) In 1963, Jackson introduced a more general Markovian model that includes the Jackson networks (1957) and the Gordon and Newell networks (1967a,b)

as particular cases. Here he considers the state-dependent external arrival rate (dependent upon the total number of customers in the system) and the state-dependent exponential-service rate (dependent on the number of customers at a node). Jackson gives sufficient conditions for the existence of equilibrium distribution at the nodes and shows that the probability distribution of the state of the network is of product form.

(3) Posner and Bernholtz (1968a,b) generalize Gordon and Newell's model by allowing, among other things, travel time between pairs of nodes to have arbitrary distribution and by providing for different service rates.

5.6 BCMP Networks

We now examine a more general network: the network considered by Baskett, Chandy, Muntz, and Palacios. It is called the BCMP network. Here new service disciplines as well as a number $R(\geq 1)$ of classes of jobs (customers) are introduced.

Let k be the number of nodes in a *closed* network and let there be $R(\geq 1)$ classes of customers (jobs). Customers circulate in the network and may change class as they move from one node to another. A job of class r and node i, after completing its service at node i, moves to node j as a job of class s with probability $p_{ir,js}$. Thus,

$$\boldsymbol{P} = (p_{ir,js}), \qquad 1 \leq i \leq k, \quad 1 \leq j \leq k, \quad 1 \leq r, \quad s \leq R$$

is the TPM of a Markov chain.

Let μ_{ir} be the service rate of a customer of class r at node i.

Nodes can be divided into four types according to the service discipline.

Type 1: The node has a single server, with exponential service time; the rates of service are the same for all R types of jobs, and the service discipline is FIFO. Here $\mu_{ir} = \mu_i$ for all r.

Type 2: The node has a single server with processor-sharing service discipline. Each class of job may have a distinct service-time distribution that is given by a differentiable function.

Type 3: The node has a sufficient number of servers so that there is no queue at the node (and so that a job starts receiving service immediately on arrival). Service-time distribution, which is to be differentiable, can be distinct for a distinct class of customers.

Type 4: The node has a single-server, preemptive LIFO service discipline such that a new arrival interrupts the customer being served, if any, and the displaced customer returns to the head of the queue and starts receiving service as soon as the customer who caused the interruption completes his service. Service-time distribution, which is to be differentiable, can be distinct for a distinct class of customers.

Service-time distribution at nodes of Types 2, 3, and 4 can be the more general Coxian type. (See Gelenbe and Pujolle (1987) for a description of Coxian distribution and its properties.) Further, distributions can be different for different classes.

We examine the system-state probability distribution in steady state.

First, solve the traffic equations for each distinct class. Let the solution be designated by α_{ir} where α_{ir} satisfies

$$\alpha_{ir} = \sum_{r'=1}^{R} \left(\sum_{j=1}^{k} \alpha_{jr} p_{jr',ir} \right), \quad i = 1, 2, \ldots, k, \quad r = 1, 2, \ldots, R.$$

Thus, α_{ir} gives the relative frequency of the number of visits of a customer of class r to the node i.

Let n_r be the number of jobs of class r in the network and n_{ir} be the number of jobs of class r at node i. Then

$$n_r = \sum_{i=1}^{k} n_{ir}, \quad r = 1, 2, \ldots, R,$$

and $m_i = \sum_{r=1}^{R} n_{ir}$ is the total number of jobs at node i, $i = 1, 2, \ldots, k$. Denote vector N_i by

$$N_i = (n_{i1}, n_{i2}, \ldots, n_{iR}), \quad i = 1, \ldots, k.$$

Then a state of the system can be denoted by the vector

$$\boldsymbol{N} = (N_1, N_2, \ldots, N_k).$$

We now state without proof the BCMP theorem.

Theorem 5.3. BCMP Theorem (Baskett *et al.*, 1975)
The steady-state probability that the system is at state $\boldsymbol{N}$ *is given by*

$$P(\boldsymbol{N}) = \frac{1}{A(n_1, \ldots, n_k)} \prod_{i=1}^{k} f_i(N_i),$$

where

$$f_i(N_i) = (m_i)! \prod_{r=1}^{R} \frac{1}{(n_{ir})!} (\alpha_{ir})^{n_{ir}} \left(\frac{1}{\mu_i} \right)^{m_i},$$

if node i *is of Type 1*

$$= (m_i)! \prod_{r=1}^{R} \frac{1}{(n_{ir})!} \left(\frac{\alpha_{ir}}{\mu_{ir}} \right)^{n_{ir}},$$

if node i *is of Type 2 or 4;*

$$= \prod_{r=1}^{R} \frac{1}{(n_{ir})!} \left(\frac{\alpha_{ir}}{\mu_{ir}} \right)^{n_{ir}}$$

if node i *is of Type 3;*

and A^{-1} *is the normalizing constant.*

Notes:

(1) The result is of product form.

(2) Only the mean service times enter in the result (even for a more general type of service distribution considered).

(3) The case of the open network can also be covered by considering two fictitious nodes: node O (origin) and node $k+1$ (sink). Then we shall have the TPM.

$$\boldsymbol{P} = (p_{ir,jr'}), \qquad \begin{aligned} i &= 0, 1, \ldots, k, \\ j &= 1, 2, \ldots, k+1, \\ r, r' &= 1, 2, \ldots, R. \end{aligned}$$

(4) The BCMP network's essential advantage is in its introduction of different classes of jobs and service disciplines other than FIFO. These considerations lead to a wider field of applications.

(5) The main difficulty in using the result of the BCMP network is the computation of the normalization constant. For computational techniques, see, for example, Reiser (1977, 1982) and Sauer and Chandy (1981).

(6) When only the FIFO service discipline is considered, then the case of only Type 1 node in BCMP network will arise. Here service time is exponential and all classes have the same service-time distribution in a node. Thus, the case can be covered by the Jackson network as well.

(7) Kelly (1975) gives a generalization of the BCMP theorem by allowing jobs to follow arbitrary paths in the network and not Bernoulli branches.

5.7 Concluding Remarks

Networks have been found to be very useful in the formulation and modeling of computer, communication, and other such systems. Several studies have been directed toward networks. Williams and Bhandiwad (1976) consider a model for multiprogrammed computers. For a general review and survey of the Jackson network, refer to Lemoine (1977, 1978), who points out the limitations of classical work and lists some open questions. For an exhaustive survey of random processes in queueing networks, see Disney and König (1985) (which contains a list of 314 references). For a review of closed network and cyclic queues, see Koenigsberg (1982). For a review of queueing-network models in computer-system design, see Kobayashi (1978), Gelenbe and Mitrani (1980), Gelenbe and Pujolle (1987), Sauer and Chandy (1981), and Geist and Trivedi

(1982). For a survey of performance evaluation of data-communication systems, see Reiser (1982) and Kobayashi (1978). See also Walrand (1988, 1990) and Kobayashi and Mark (1997).

Closed Markovian queueing networks have emerged as an important tool for modeling computer systems, communication systems, online computer networks, and other real-time computer-based systems. This has been possible because of the discovery of an important class of networks, the so-called product-form networks, which are analytically tractable (Sauer and Chandy 1981; Kelly, 1979). In addition to networks, discrete-time queueing processes provide suitable models for analysis of data traffic in computer and communication systems. (See Kobayashi (1983) for discrete-time queueing systems.)

Progress in the theory of Markovian networks has resulted in several algorithms for efficient computation of performance measures such as utilization, throughput, average response time, and marginal distributions. The convolution method has been treated by Buzen (1973) and Reiser and Kobayashi (1975). Kobayashi (1978, 1983) proposes Polyatheoretic algorithms.

The design and optimal operation in routing queueing networks have also been of considerable importance. Such questions arise in management of computer systems and in other areas. The problem relates to finding appropriate parameter values for optimization of a specified objective function. (See, for example, Lazar (1982, 1983) and Schwartz (1977). Algorithmic results for determination of optimal routing strategies in a network have been given, for example, by Agnew (1976), Gallager (1977), and Towsley (1980).

Sojourn time in queueing networks has also been engaging considerable attention. Refer to papers, for example, by Boxma and Donk (1982), Boxma *et al.* (1984), Chow (1980), Daduna (1986a,b), Kelly and Pollett (1983), Lemoine (1979, 1987), Melamed (1982), Schassberger and Daduna (1983, 1987), Walrand and Varaiya (1980), and Balsamo and Donatiello (1989).

The literature on queueing networks has been growing at a very rapid pace. For some details and for references, see Gelenbe and Pujolle (1987) and Walrand (1988, 1990).

5.7.1 Loss networks

There is another kind of network, known as a loss network, that has several analogous properties with queueing networks. Similar methods apply to both these networks. Loss networks are also known as circuit-switched networks, since these provide accurate models for such networks.

A Jackson network (which is a queueing network) may be thought of as a generalization of the $M/M/1$ queue, where several queues form a network. A loss network is a generalization of the classical Erlang Loss model where multiple customer and server classes are introduced and are allowed to have multiple servers simultaneously. The theory of loss networks has, of late, been of great

interest in view of its application to the design and control of telecommunication systems.

Loss networks having product form and insensitivity properties of Jackson networks have been receiving wide attention. For a discussion of loss networks, refer to Kelly (1991), Kobayashi and Mark (1997), and the references therein. See also Bertsekas and Gallager (1992) for data networks.

Problems and Complements

5.1. Consider an open network with two nodes having a single exponential server at each of two nodes with service rates $\mu_i, i = 1,2$. Suppose that arrivals to node 1 occur in accordance with a Poisson process having rate λ. After being served at node 1, the customer (job) goes to node 2 with probability p or leaves the system with probability $(1 - p) = q$. From node 2 it again goes to node 1. Assume that the system is at steady state. Show that the queues at nodes 1 and 2 behave like independent $M/M/1$ queues with intensities

$$\rho_1 = \frac{\lambda}{(q\mu_1)} \quad \text{and}$$

$$\rho_2 = \frac{\lambda p}{(q\mu_2)},$$

and that the system state (n_1, n_2) with n_i at node $i, i = 1,2$, has the probability

$$p(n_1, n_2) = (1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2}.$$

Show that the average response time at two nodes is given by

$$E(R) = \frac{1}{\lambda}\left[\frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{1 - \rho_2}\right].$$

Let $E(B_i)$ be the average service-time requirements at node $i, i = 1,2$, Then show that

$$E(R) = \sum_{i=1}^{2} \frac{E(B_i)}{1 - \lambda E(B_i)}.$$

5.2. Consider an open network such that arrivals to node 0 occur from outside in accordance with a Poisson process with rate λ. After receiving service at node 0, the job (customer) may leave the system with probability q or may go one of the k nodes, the probability that it goes to node i being $p_i, i = 1, \ldots, k, \sum_{i=1}^{k} p_i = 1 - q$. From node $i, i = 1,2, \ldots, k$, it goes

back to node 0. Each of the $(k+1)$ nodes has an exponential server, the rate at node i being $\mu_i, i = 0, 1, 2, \ldots, k$.

Represent the network by a suitable diagram. Write down the routing matrix $\boldsymbol{P}$ and find $\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \ldots, \alpha_k)$. Assume that the system is at steady state. Show that the queues at the $(k+1)$ nodes behave as independent $M/M/1$ queues. Denoting

$$\rho_i = \frac{\alpha_i}{\mu_i}, \quad i = 0, 1, 2, \ldots, k,$$

show that the system state $(n_0, n_1, \ldots, n_k)$ has probability

$$p(n_0, n_1, \ldots, n_k) = \prod_{i=0}^{k} (1 - \rho_i)\rho_i^{n_i}.$$

Show that the average total response time is given by

$$E(R) = \frac{1}{\lambda} \sum_{i=0}^{k} \frac{\rho_i}{1 - \rho_i}$$

$$= \sum_{i=0}^{k} \frac{E(B_i)}{1 - \lambda E(B_i)}$$

where $E(B_i)$ is the average service-time requirement at node i (Trivedi, 1982).

5.3. $M/M/1$ Queue with Bernoulli feedback.

Consider an $M/M/1$ FCFS queue with Bernoulli feedback such that after completion of service, the job may leave the system with probability q or may be fed back into the system with probability p.

(a) Show that the effective average arrival rate to the queue is λ/q.

(b) Show that the number of jobs N in the system has a geometric distribution

$$Pr\{N = n\} = \left(1 - \frac{\lambda}{q\mu}\right)\left(\frac{\lambda}{q\mu}\right)^n, \quad n = 0, 1, 2, \ldots.$$

(c) Show that the time Y from the last input (to the server) to the next feedback has the distribution

$$R_Y(t) = Pr\{Y \geq t\} = \frac{p\mu}{\mu - \lambda} e^{-(\mu - \lambda)t} + \frac{q\mu - \lambda}{\mu - \lambda}$$

and that the interinput time I has hyperexponential distribution.

(d) Show that the number of jobs N_d left behind by a customer departing from the system has the same distribution as N—that is,

$$Pr\{N_d = n\} = Pr\{N = n\}, \quad n = 0, 1, 2, \ldots.$$

(e) Further, show that the departure process is Poisson with rate λ (Burke, 1976).

5.4. Consider Jackson's open-network model with a single server at each of the k nodes. Let T and S denote, respectively, the total time spent in the system and the total service time received by a unit. Assume that the system is in equilibrium and that the network is such that a unit can never visit any node more than once. If μ_i is the rate of service for the server (exponential) at node i and $\rho_i = \alpha_i/\mu_i, i = 1, 2, \ldots, k$, then show that

$$E(T) = A^{-1} \sum_{i=1}^{k} \frac{\rho_i}{1 - \rho_i}$$

$$E(S) = A^{-1} \sum_{i=1}^{k} (1 + \rho_i)$$

and

$$E(W) = E(T - S) = A^{-1} \sum_{i=1}^{k} \frac{\rho_i^2}{1 - \rho_i^2},$$

where $A = \sum_{i=1}^{k} \lambda_i, \lambda_i$ being the rate of arrival from outside to node i. Show that the LST of T is the transform of a mixture of exponential distributions and find the same (Lemoine, 1977).

5.5. Consider a closed network with three nodes and K circulating jobs. Suppose that the service times at the nodes are independent exponential RVs with rates μ_1, μ_2, and μ_3, respectively. Suppose that after receipt of service at node 1, the job is fed back into node 1 with probability p_1 or goes to nodes 2 (or 3) with probability p_2 (or p_3), with ($p_1 + p_2 + p_3 = 1$). Draw a diagram and find $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. Show that the probability that the system is at state $(n_1, n_2, n_3), \sum_{i=1}^{3} n_i = K$ is given by

$$p(n_1, n_2, n_3) = \frac{1}{A(K)} \left(\frac{p_2\mu_1}{\mu_2}\right)^{n_2} \left(\frac{p_3\mu_1}{\mu_3}\right)^{n_3}$$

where

$$A(K) = \sum_{\substack{n_i \geq 0 \\ \sum n_i = K}} \prod_{i=0}^{k} \left(\frac{p_2\mu_1}{\mu_2}\right)^{n_2} \left(\frac{p_3\mu_1}{\mu_3}\right)^{n_3}.$$

Show that the server utilization at node 1 equals

$$u_1 = \sum_{\substack{n_i \geq 0 \\ \sum n_i = K}} \prod_{i=0}^{k} p(n_1, n_2, n_3) = \frac{A(K-1)}{A(K)}.$$

5.6. $M/G/1/K$ queue as Cyclic Network. (Refer to the next chapter for an analysis of $M/G/1$ queues.)
Consider a cyclic network with K circulating jobs and two nodes, with a single server at each of them. The server at node 1 has a general service-time distribution (G) with rate μ, and the server at node 2 has an exponential service-time distribution (M) with rate λ. Suppose that the system is in steady state. Let

$$p(n, K-n) = Pr\{n \text{ units at node 1 and } K-n \text{ units at node 2 in the cyclic network}\},$$

$$p_k(n) = Pr(n \text{ unit in the system in an } M/G/1/K \text{ limited-waiting-space queue}),$$

and $$p_n = Pr\{n \text{ units in an unrestricted } M/G/1/\infty \text{ queue}\}.$$

(a) Show that $p_k(n) = p(n, K-n), n = 0, 1, 2, \ldots, K$ (that is, the cyclic queue described previously is equivalent to an $M/G/1/K$ queue).
(b) Show further that, for $\rho = \lambda/\mu$,

$$p_K(n) = C(K)p(n), \quad n = 0, 1, \ldots, K-1$$
$$= 1 - \frac{\{1 - C(K)(1-\rho)\}}{\rho}, \quad n = K$$

where

$$C(K) = \left\{1 - \rho\left[1 - \sum_{n=0}^{K-1} p(n)\right]\right\}^{-1}.$$

(Gelenbe and Pujolle, 1987)
(c) Show that $u_1 \equiv 1 - p(0, K), u_2 = 1 - p(K, 0)$ are the utilization factors at nodes 1 and 2 and that $u_1/u_2 = \rho$.
Verify the results for the particular case $G \equiv M$.

Note: Carroll *et al.* (1982) consider such a two-node loop system by taking a Coxian server at one node. The results are extended by considering two Coxian servers at the two nodes by Van de Liefvoort (1986).

5.7. $M/G/1$-PS model
Consider an $M/G/1$ queue where the queue discipline is processor sharing or time-sharing. This discipline implies that if there are already $(n-1)$

customers in the system, then the arriving customer as well as the other (waiting) customers in the system all start receiving service immediately at the average rate of μ/n. There is no queue as such, and the rate at which units receive service changes each time a new arrival joins the system and each time a unit whose service requirement is fully met departs from the system.

We denote the system by $M/G/1$-PS.

(a) Show that the steady-state distribution of the number in the system N has the same geometric distribution as that in a standard $M/M/1$ queue (with FIFO discipline)—that is, for $\rho = \lambda/\mu < 1$,

$$Pr\{N = n\} = (1 - \rho)\rho^n, \quad n = 0, 1, 2, \ldots.$$

(b) If B is DF of the service-time distribution S with mean $E(S)$ and W is the response time (waiting time in the system), then show that

$$E\{W \mid S = t\} = \frac{t}{1 - \rho}$$

and that

$$E\{W\} = \int_0^\infty \frac{t}{1 - \rho}\, dB(t) = \frac{E(S)}{1 - \rho}.$$

(c) Show that the average conditional delay (see Note (1) below) experienced by a customer is given by

$$\begin{aligned} E[D \mid S = t] &= E[W \mid S = t] - t \\ &= \frac{\rho t}{1 - \rho} \end{aligned}$$

and that

$$E[D] = \frac{\rho E(S)}{1 - \rho}.$$

(d) Show further that the output process is Poisson (Kleinrock, 1967).

Notes:

(1) A customer experiences a delay D as the full service time required by the customer cannot be had in a single installment because the discipline is processor-sharing (though there is no queue as such and an arrival starts receiving service immediately on arrival).

(2) For an $M/G/1$-PS queue, $E[W]$ depends on the distribution of the service time S only through its expected value $E[S]$, whereas in case of a standard $M/G/1$-FIFO service, both the first and second moments of S enter in the expression for $E[W]$ (as is given by the Pollaczek-Khinchine formula).

(3) $M/G/1$-PS can be used as model of some time-shared computer systems.

(4) The distribution of D is not known in the general case. See the next problem for the special case $G \equiv M$.

5.8. $M/M/1$-PS: conditional delay

Let $F(x; t) = Pr\{D \leq x \mid S = t\}$ be the DF of the conditional delay, given that total service requirement is of length t, and let $F^*(s; t)$ be its LST. Then, when $\rho < 1$,

$$F^*(s; t) = \frac{(1-\rho)(1-\rho r^2)\exp\{-\lambda(1-r)t\}}{\{(1-\rho r^2\} - \rho(1-r)^2 \exp\left\{\frac{-\mu(1-\rho r^2)t}{r}\right\}},$$

where r is the (smaller) root of

$$\lambda x^2 - (\lambda + \mu + s)x + \mu = 0.$$

Deduce that

$$E\{D \mid S = t\} = \frac{\rho t}{1-\rho} \quad \text{and}$$

$$var\{D \mid S = t\} = \left[\frac{2\rho t}{\mu(1-\rho)^3} - \frac{2\rho}{\mu(1-\rho)^4}\right] \times [1 - \exp\{-(1-\rho)\mu t\}].$$

Compare $E\{D \mid S = t\}$ for $M/M/1$-PS with average queueing time $E\{W_q\}$ for the $M/M/1$-FIFO discipline.

Show that for $t < 1/\mu$,

$$E\{D \mid S = t\} < E\{W_q\},$$

that is, for arrivals whose service-time requirement t is less than the average service time $1/\mu$, the mean delay under the PS discipline is less than the mean delay (queueing time) under the FIFO discipline.

Show that, for $\rho < 1$,

$$E\{D \mid S = t, N = n\} = \frac{\rho t}{1-\rho} + \frac{[n(1-\rho) - \rho)][1 - \exp\{-(1-\rho)\rho t\}]}{\mu(1-\rho)^2}$$

where N is the number in the system in an $M/M/1$-PS system.

Verify that

$$\sum_{n=0}^{\infty} E\{D \mid S = t, N = n\} Pr\{N = n\} = E\{D \mid S = t\}.$$

Show further that

$$\lim_{\rho \to 1} E\{D \mid S = t, N = n\} = nt + \frac{\mu t^2}{2}.$$

(Coffman *et al.*, 1970)

5.9. Sojourn-time distribution in cyclic exponential network.
Consider a cyclic network with K circulating jobs among two nodes: 1 and 2. The service time at node i is exponential with parameter $\mu_i, i = 1, 2$, and the service processes at the two nodes are independent. Let W_i be the sojourn time or response time (queueing plus service time) at node $i, i = 1, 2$. Show that the joint distribution of the consecutive sojourn times W_1, W_2 has the LST

$$E[\exp(-s_1 W_1 - s_2 W_2)]$$

$$= \sum_{n=0}^{K-1} a(n) \left(\frac{1}{1 + \frac{s_1}{\mu_1}} \right)^{n+1} \left(\frac{1}{1 + \frac{s_2}{\mu_2}} \right)^{K-n}, \quad \text{Re } s_1, s_2 \geq 0.$$

where

$$a(n) = \frac{1 - \left(\frac{\mu_2}{\mu_1}\right)}{1 - \left(\frac{\mu_2}{\mu_1}\right)^K} \left(\frac{\mu_2}{\mu_1} \right)^n, \qquad \mu_1 \neq \mu_2,$$

$$= \frac{1}{K}, \qquad \mu_1 = \mu_2.$$

Show that the correlation between W_1 and W_2 is nonpositive (Boxma and Donk, 1982).

Notes:

(1) Boxma *et al.* (1984) extend (by induction) the result to an $M(\geq 1)$ node exponential cyclic network.

(2) The distribution has a product form.

5.10. Cycle-time distribution in cyclic exponential network.
Consider an M-node cyclic network with K circulating jobs, the nodes having independent exponential servers with parameters $\mu_i, i = 1, 2, \ldots, M$. Let W_i be the sojourn time at node i and let

$$T = W_1 + \cdots + W_M$$

be the cyclic time (sum of consecutive sojourn times). Show that the LST of T is given by

$$E[\exp(-sT)] = E\{\exp[-s(W_1 + \cdots + W_M)]\}$$

$$= \sum_{n_i \in C} a(n_1 \ldots, n_M) \prod_{i=1}^{M} \left(\frac{1}{1 + s/\mu_i} \right)^{n_i + 1} \quad \text{Re } s \geq 0,$$

where

$$C \equiv C(M, K-1) = \left\{ (n_1, \ldots, n_M) : n_i \geq 0, \sum_{i=1}^{M} n_i = K-1 \right\} \text{ and}$$

$$a(n_1, \ldots, n_M) = \prod_{i=1}^{M} \frac{\left(\frac{1}{\mu_i}\right)^{n_i}}{\sum_{n_i \in C} \prod_{i=1}^{M} \left\{\frac{1}{\mu_i}\right\}^{n_i}}.$$

Find $E(T)$ for $M = 2$ (Schassberger and Daduna, 1983). Note that the distribution has a product form.

References and Further Reading

Agnew, C. E. (1976). On quadratic adaptive routing algorithms. *Com. Ass. Comp. Mach.* **19**, 18–22.

Allen, A. O. (1990). *Probability, Statistics and Queueing Theory with Computer Science Applications*, 2nd ed., Academic Press, New York.

Balsamo, S., and Donatiello, L. (1989). On the cycle time distribution in a two-stage cyclic network with blocking. *IEEE Trans. Software Eng.* **15**, 1206–1216.

Barbour, A. D. (1976). Networks of queues and the method of stages. *Adv. Appl. Prob.* **8**, 584–591.

Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. (1975). Open, closed and mixed networks of queues with different classes of customers. *J. Ass. Comp. Mach.* **22**, 248–260.

Bertsekas, D., and Gallager, R. (1992). *Data Networks*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Boxma, O. J. (1983). The cyclic queue with one general and one exponential server. *Adv. Appl. Prob.* **15**, 857–873.

Boxma, O. J. (1986). Models of Two Queues: a Few New Views in *Teletraffic Analysis and Computer Performance Evaluation*, 75–98 (Eds. O. J. Boxma, J. W. Cohen, and H. C. Tijms), Elsevier Science Publishers B. V., North Holland, Amsterdam. (The paper contains a list of 94 references.)

Boxma, O. J., and Donk, P. (1982). On response time and cycle time distributions in a two-stage cyclic queue. *Perf. Ev.* **2**, 181–194.

Boxma, O. J., Kelly, F. P., and Konheim, A. G. (1984). The product form for sojourn time distributions in cyclic exponential queues. *J. Ass. Comp. March.* **31**, 128–133.

Brandwajn, A., and Jow, Y. L. (1988). An approximation method for tandem queues with blocking. *Opns. Res.* **36**, 73–88.

Burke, P. J. (1956). The output of a queueing system. *Opns. Res.* **4**, 699–704.

Burke, P. J. (1976). Proof of a conjecture on the interarrival time distribution in an $M/M/1$ queue with feedback. *IEEE Trans. On Comm.* **24**, 175–178.

Buzacott, J. A., and Yao, D. D. (1986). On queueing network models of flexible manufacturing systems. *Queueing Systems* **1**, 5–27.

Buzen, J. P. (1973). Computational algorithms for closed queueing networks with exponential servers. *Comm. Ass. Comp. Mach.* **16**, 527–531.

Carroll, J. K., Van de Liefvoort, A., and Lipsky, L. (1982). Solutions of $M/G/1/N$ type loops with extensions to $M/G/1$ and $GI/M/1$ queues. *Opns. Res.* **30**, 490–514.

Chandy, K., Herzog, U., and Woo, L. (1975a). Parametric analysis of queueing networks. *IBM J. Res. Dev.* **19**, 36–42.

Chandy, K. M., Herzog, U., and Woo, L. (1975b). Approximate analysis of general queueing networks. *IBM J. Res. Dev.* **19**, 43–49.

Chandy, K. M., Howard, J. H., and Towsley, D. F. (1977). Product form and local balance in queueing networks. *J. Ass. Comp. Mach.* **24**, 250–263.

Chen, T. M. (1989). On the independence of sojourn times in tandem queues. *Adv. Appl. Prob.* **21**, 488–489.

Chow, W. M. (1980). The cycle time distribution of exponential queues. *J. Ass. Comp. Mach.* **27**, 281–286.

Coffman, E. G., Jr., Muntz, R. R., and Trotter, H. (1970). Waiting time distributions for processor sharing systems. *J. Ass. Comp. Mach.* **17**, 123–130.

Daduna, H. (1986a). Cycle times in two-stage closed queueing network: applications to multi-programmed computer systems with virtual memory. *Opns. Res.* **34**, 281–288.

Daduna, H. (1986b). Two-stage cycle queues with nonexponential servers: steady state and cyclic time. *Opns. Res.* **34**, 455–459.

Disney, R. L. (1981). Queueing networks. *American Math. Soc. Proceedings of Symposia in Applied Mathematics* **25**, 53–83.

Disney, R. L., and König, D. (1985). Queueing networks. A survey of their random processes. *SIAM Review* **27**, 335–403. (Contains a list of 314 references.)

Disney, R. L., and Kiessler, P. C. (1987). *Traffic Processes in Queueing Networks: A Markov Renewal Approach*, The Johns Hopkins University Press, Baltimore, MD.

Dukhovny, I. M., and Koenigsberg, E. (1981). Invariance properties of queueing networks and their application to computer communication systems. *INFOR* **19**, 185–204.

Gallager, R. R. (1977). A minimum delay routing algorithm using distributed component. *IEEE Trans. on Comm.* **25**, 73–84.

Geist, R., and Trivedi, K. (1982). Queueing network models in computer system design. *Mathematics Magazine* **55**, 2, 67–80.

Gelenbe, E. (1975). On approximate computer system models. *J. Ass. Comp. Mach.* **22**, 261–269.

Gelenbe, E., and Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems*, Academic Press, New York.

Gelenbe, E., and Pujolle, G. (1987). *Introduction to Queueing Networks*, (2nd ed. 1998) John Wiley, New York.

Gnedenko, B. V., and König, D. (Eds.) (1983). *Handbuch der Bedienungs-Theorie, II. Formeln andere Ergebnisse,* Academic-Verlag, Berlin.

Gnedenko, B. V., and Kovalenko, I. N. (1968). *Introduction to Queueing Theory* (2nd ed. 1980), Israel Program for Sci. Tran, Jerusalem.

Goodman, J. B., and Massey, W. A. (1984). The non-ergodic Jackson network. *J. Appl. Prob.* **21**, 860–869.

Gordon, W. J., and Newell, G. P. (1967a). Closed queueing systems with exponential servers. *Opns. Res.* **15**, 254–265.

Gordon, W. J., and Newell, G. P. (1967b). Cyclic queueing systems with restricted length queues. *Opns. Res.* **15**, 266–277.

Gross, D., and Harris, C. M. (1985). *Fundamentals of Queueing Theory*, 2nd ed. (3rd ed. 1998), Wiley, New York.

Heidelberger, P., and Lavenberg, S. (1984). Computer performance evaluation methodology. *IEEE Trans. on Computers* **C-33**, 1195–1220.

Jackson, J. R. (1957). Networks of waiting lines. *Opns. Res.* **5**, 518–522.

Jackson, J. R. (1963). Jobshop like queueing systems. *Mgmt. Sci.* **10**, 131–142.

Jackson, R. R. P. (1954). Queueing systems with phase type service. *Opl. Res. Q.* **5**, 109–120.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis*, J. Wiley & Sons, New York.

Jansen, V., and König, D. (1980). Insensitivity and steady state probabilities in product form for queueing networks. *J. Inf. Proc. & Cyber (EIK)* **16**, 385–397.

Kelly, F. P. (1975). Networks of queues with customers of different types. *J. Appl. Prob.* **12**, 542–554.

Kelly, F. P. (1976). Networks of queues. *Adv. Appl. Prob.* **8**, 416–432.
Kelly, F. P. (1979). *Reversibility and Stochastic Networks*, Wiley, New York.
Kelly, F. P. (1985). Stochastic models of computer communication systems. *J.R.S.S.* **B47**, 379–395.
Kelly, F. P. (1989). On a class of approximation for closed queueing networks. *Queueing Systems* **4**, 69–76.
Kelly, F. P. (1991). Loss networks. *Annal. Appl. Prob.* **1**, 319–378. (Contains a list of 83 references.)
Kelly, F. P., and Pollett, P. K. (1983). Sojourn times in closed queueing networks. *Adv. Appl. Prob.* **15**, 638–656.
Kiessler, P. C. (1989). A simple proof of the equivalence of input and output intervals in Jackson networks. *Opns. Res.* **37**, 645–647.
Kleinrock, L. (1967). Time shared systems—a theoretical treatment. *J. Ass. Comp. Mach.* **14**, 242–261.
Kobayashi, H. (1978). *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*, Addison-Wesley, Reading, MA.
Kobayashi, H. (1983). Stochastic Modeling: Queueing Models in *Probability Theory and Computer Science* (Eds. G. Louchard and G. Latouche), Academic Press, London.
Kobayashi, H., and Konheim, A. G. (1977). Queueing models for computer communication system analysis. *IEEE Trans. Comm.* **COM-25**, 2–29.
Kobayashi, H., and Mark, B. L. (1994). On queueing networks and loss networks. *Proc. 28th Annual Conf. on Info. Sci & Sys.*, 794–799.
Kobayashi, H., and Mark, B. L. (1997). Product-Form Loss Networks in *Frontiers in Queueing* (Ed. J. H. Dshalalow) 147–195, Boca Raton, FL. (Contains a list of 43 references.)
Kobayashi, H., and Reiser, M. (1975). On generalization of job routing behavior in a queueing network model. *IBM Research Report*, **RC** 5252.
Koenigsberg, E. (1982). Twenty-five years of cycle queues and closed queue networks: a review. *J. Opl. Res. Soc.* **33**, 605–619.
Koenigsberg, E. (1985). Cyclic queues. *Opns. Res. Qrly*, **9**, 22–35.
Koenigsberg, E. (1983). Dominance, Invariance, conservation, decomposition, and equivalence applied to queue networks. Preprint.
Lavenberg, S. S. (Ed.) (1983). *Computer Performance Modeling Handbook.* Academic Press, New York.
Lazar, A. A. (1982). The throughput time delay function of an $M/M/1$ queue. *IEEE Trans. Inf. Th.* **29**, 914–918.
Lazar, A. A. (1983). Optimal flow control of a class of queueing networks in equilibrium. *IEEE Trans. Aut. Cont*, **AC-28**, 1001–1007.
Lemoine, A. J. (1977). Network of queues—a survey of equilibrium analysis. *Mgmt. Sci.* **24**, 464–481.
Lemoine, A. J. (1978). Networks of queues—a survey of weak convergence results. *Mgmt. Sci.* **24**, 1175–1193.
Lemoine, A. J. (1979). On total sojourn time in networks of queues. *Mgmt. Sci.* **25**, 1034–1035.
Lemoine, A. J. (1987). On sojourn time in Jackson networks of queues. *J. Appl. Prob.* **24**, 495–510.
Massey, W. (1984a). Open networks of queues: their algebraic structure and estimating their transient behavior. *Adv. Appl. Prob.* **16**, 176–201.
Massey, W. (1984b). An operator-analytic approach to Jackson network. *J. Appl. Prob.* **21**, 379–393.
McKenna, J. (1989). A generalization of Little's law to moments of queue length and waiting times in a closed product form queueing network. *J. Appl. Prob.* **26**, 121–133.
Melamed, B. (1979). Characterization of Poisson traffic streams in Jackson queueing networks. *Adv. Appl. Prob.* **11**, 422–438.

Melamed, B. (1982). Sojourn times in queueing networks. *Math. Opns. Res.* **7**, 233–244.
Molloy, M. (1989). *Fundamentals of Performance Modeling*, Macmillan, New York.
Newell, G. F. (1984). Approximations for superposition arrival processes in queues. *Mgmt. Sci.* **30**, 623–630.
Perros, H. G. (1984). Queueing networks with blocking: a bibliography. *Acta Sigmetrics* **10**, 139–148.
Perros, H. G., and Altiok, T. (Eds.) (1989). *Queueing Networks with Blocking*, North Holland, New York.
Posner, M., and Bernholtz, B. (1968a). Closed finite queueing networks with time lags. *Opns. Res.* **16**, 962–976.
Posner, M., and Bernholtz, B. (1968b). Closed finite queueing networks with time lags and several classes of units. *Opns. Res.* **16**, 977–985.
Reiser, M. (1977). Numerical Methods in Separable Queueing Networks in *Algorithmic Methods in Probability* (Ed. M. F. Neuts), TIMS *Studies in Management Sciences*, Vol. 7, 113–142. North-Holland, Amsterdam.
Reiser, M. (1982). Performance evaluation of data communication systems. *Proc. of the IEEE* **70**, 171–196. (Contains a list of 142 references.)
Resier, M., and Kobayashi, H. (1975). Queueing networks with multiple closed chains: theory and computational algorithms. *IBM J. Res. Develop.* **19**, 282–294.
Resier, M., and Lavenberg, S. S. (1980). Mean value analysis of closed multichain queueing networks. *J. Ass. Comp. Mach.* **27**, 313–322.
Robertazzi, T. G. (1990). *Computer Networks and Systems: Queueing Theory and Performance Evaluation*, Springer-Verlag, New York.
Sauer, C. H., and Chandy, K. (1981). *Computer System and Performance Modeling*, Prentice-Hall, Englewood Cliffs, NJ.
Schassberger, R., and Daduna, H. (1983). The time for a round trip in a cycle of exponential queues. *J. Ass. Comp. Mach.* **30**, 146–150.
Schassberger, R., and Daduna, H. (1987). Sojourn times in queueing networks with multi-server nodes. *J. Appl. Prob.* **24**, 511–421.
Schwartz, M. (1977). *Computer Communication Network Designs and Analysis*, Prentice-Hall, Englewood Cliffs, NJ.
Schwartz, M. (1987). *Telecommunication Networks: Protocols, Modeling and Analysis*, Addison-Wesley, Reading, MA.
Shalmon, M., and Kaplan, M. (1984). A tandem network of queues with deterministic service and intermediate arrivals. *Opns. Res.* **32**, 753–773.
Shanthikumar, J. G., and Yao, D. D. (1988a). Stochastic monotonicity of the queue length in closed queueing networks. *Opns. Res.* **35**, 583–588.
Shanthikumar, J. G., and Yao, D. D. (1988b). Second order properties of the throughput of a closed queueing network. *Maths. Opns. Res.* **13**, 524–534.
Shanthikumar, J. G., and Yao, D. D. (1989). Stochastic monotonicity in general queueing networks. *J. Appl. Prob.* **26**, 413–417.
Sigman, K. (1989). Notes on stability of closed queueing networks. *J. Appl. Prob.* **26**, 678–682.
Towsley, D. (1980). Queueing network models with state-dependent routing. *J. Ass. Comp. Mach.* **27**, 323–337.
Trivedi, K. S. (1982). *Probability and Statistics with Reliability, Queueing and Computer Science Applications*, (2nd ed. 2001) Prentice-Hall, Englewood Cliffs, NJ.
Van de Liefvoort, A. (1986). A matrix-algebraic solution to two K_m servers in a loop. *J. Ass. Comp. Mach.* **33**, 207–223.
Walrand, J. (1982a). On the equivalence of flows in a network of queues. *J. Appl. Prob.* **19**, 195–203.
Walrand, J. (1982b). Poisson flows in single class open networks of quasi reversible queues, *Stoch. Proc. & Appl.* **13**, 293–303.
Walrand, J. (1988). *Introduction to Queueing Networks*, Prentice-Hall, Englewood Cliffs, NJ.

Walrand, J. (1990). Queueing Networks in *Handbooks in Operations Research and Management* (Eds. D. P. Heyman and M. J. Sobel), Vol. 2, pp. 519–604, North-Holland, Amsterdam.

Walrand, J., and Varaiya, P. (1980). Sojourn times and the overtaking condition in Jacksonian networks. *Adv. Appl. Prob.* **12**, 1000–1018.

Walrand, J., and Varaiya, P. (1981). Flows in queueing networks: a martingale approach. *Math. of Opns. Res.* **6**, 387–404.

Weber, R. R. (1979). The interchangeability of tandem queues. *J. Appl. Prob.* **16**, 690–695.

Whitt, W. (1983). The queueing network analyzer. *Bell. Sys. Tech. J.* **62**, 2779–2815.

Whitt, W. (1984). Open and closed models of network of queues. *Bell. Sys. Tech. J.* **63**, 1911–1979.

Whittle, P. (1986). *Systems in Stochastic Equilibrium*, Wiley, New York.

Williams, A., and Bhandiwad, R. (1976). A generating function approach to queueing network analysis of multiprogrammed computers. *Network* **6**, 1–22.

Yao, D. D., and Kim, S. C. (1987). Some order relations in closed network of queues with multiserver stations. *Nav. Res. Log. Qrly.* **34**, 53–66.

This Page Intentionally Left Blank

Non-Markovian Queueing Systems

6.1 Introduction

We have so far been discussing queueing processes that are either birth-death or non-birth-death processes—in either case the processes being Markovian. The theory of Markov processes could be applied in studying them. The models in which both the interarrival-time and the service-time distributions are exponential are birth-death Markovian. Models in which the distributions of either or both are Erlangian are non-birth-death but nevertheless can be treated as Markovian. We shall now consider where the distributions of the interarrival and service times are not necessarily exponential or Erlangian. The distributions are assumed to be independent.

We shall first examine a model of the type $M/G/1$, where G denotes that the distribution of service time is general.

Whereas one is often justified in considering an arrival process as Poisson (because of the superposition theorem), it seems that there is no such justification that the service time would also have the memoryless property (that the requirement of service at a point of time is independent of the amount of service time received to that point).

Nevertheless, we shall see that in some significant cases, the performance measures are insensitive to the form of the service-time distribution. Further, some data also show that assumption of exponential service time gives sufficiently accurate results.

The process $\{N(t), t \geq 0\}$, where $N(t)$ gives the state of the system or the system size at time t, is then non-Markovian. However, the analysis of such a process could be based on a Markovian process that can be extracted out

of it. There are a number of techniques or approaches that are used for this purpose.

(1) *Embedded-Markov-chain technique.* Kendall (1951) uses the concept of regeneration point (due to Palm (1943)) by suitable choice of regeneration points that involves extraction from the process $\{N(t), t \geq 0\}$ Markov chains in discrete time at those points. The technique is known as the embedded-Markov-chain technique.

(2) *The supplementary-variable technique.* This involves inclusion of such variables—for example, the service time $X(t)$ already received by the customer in service, if any. The method is based on the treatment of the couplet $\{N(t), X(t)\}$. Cox (1955), Kendall (1953), and Keilson and Kooharian (1960) have indicated this technique. Henderson (1972) considers the couplet $\{N(t), Y(t)\}$, where $Y(t)$ is the remaining or residual service time of the customer in service, if any. These methods involve inclusion of a supplementary variable $X(t)$ or $Y(t)$.

(3) *Lindley's integral-equation method.* This method, which is suitable for a $G/G/1$ system, takes the customer-arrival times as regeneration points and considers the waiting time of the nth customer (which is a Markov process) as the object of study.

(4) *Other methods.* Besides the preceding, there are other methods of dealing with non-Markovian systems such as the random-walk and combinatorial approaches (Takács, 1967) and the method of Green's function (Keilson, 1965).

Next we shall discuss the embedded-Markov-chain technique.

6.2 Embedded-Markov-Chain Technique for the System with Poisson Input

We are concerned at any instant t with a pair of RVs $N(t)$, the number in the system at time t, and $X(t)$, the service time already received by the customer in service, if any. While $\{N(t), t \geq 0\}$ is non-Markovian, the vector $\{N(t), X(t), t \geq 0\}$ is a Markov process. Whereas in the case of an $M/M/1$ system (because of the memoryless property of service-time distribution), attention can be confined to $N(t)$ alone, for the system $M/G/1$ we have to consider $X(t)$ also along with $N(t)$. Now by observing the number in the system at a select set of points rather than at all points of time t, it is possible to simplify matters to a great extent. These special sets of points or instants should be such that by considering the number in the system at any such point and other inputs, it should be possible to calculate the number in the system

at the next such point or instant. There are several such sets of points. A very suitable set of points is the set of *departure instant* (from service channel) at which successive customers leave the system on completion of service. Let the departure instants of the customers $C_1, C_2, \ldots, C_n, \ldots$ be $t_1, t_2, \ldots, t_n \ldots$, respectively. At such a point of time—say, the departure instant t_n of C_n—the time spent in service by the next customer C_{n+1} is zero and, thus, given $N(t_n)$ at any departure instant (that is, the number of customers left behind by the departing customer C_n) and given the additional input to the system (arrivals during the time of service of the next customer C_{n+1}), it is possible to calculate $N(t_{n+1})$, the number left behind by the next departing customer C_{n+1}. Thus, we get $N(t_{n+1})$ given $N(t_n)$, and the number of arrivals during the service time of customer C_{n+1}. So $\{N(t_n), n \geq 1\}$ defines a Markov chain, the instants $t_1, t_2, \ldots, t_n$ being embedded Markovian points. Thus, we can get $N(t_n)$ and its distribution— that is, the distribution of the number in the system at departure epochs $t_n, n \geq 1$.

For a queueing system (in steady state) with Poisson arrivals, we have the following properties.

(1) The probability a_n of the number n found by an arriving customer is equal to the probability d_n of the number n left behind by a departing customer. Again, Poisson arrivals see time averages. When equilibrium is reached in a queueing system with Poisson arrivals, we have $a_n = p_n$, where p_n is the probability that the number in the system at any time (in steady state) is n. Thus,

$$a_n = d_n = p_n.$$

Thus, the probability distribution of the number in the system at the embedded Markov points is the same as the probability distribution of the number in the system at all points of time. Thus, it suffices to consider the process $\{N(t_n), n \geq 0\}$ at the departure instants or, to be more specific, the process $\{N(t_n + 0), n \geq 0\}$, where $N(t_n + 0)$ is the number immediately following the nth departure. (See also Remark, Section 6.3.3.)

(2) The transitions of the process occur at departure instant t_n. The number in the system immediately following these instants form a Markov chain such that the transitions occur at the departure instants. The interval between two transitions (that is, between two departures) is equal to the service time when the departure leaves at least one in the system and is equal to the convolution of the interarrival time (which is exponential) and the service time when the departure leaves the system empty. If $Y(t)$ denotes the number of customers left behind by the most recent departing customer—that is, $Y(t) = N(t_n), t_n \leq t \leq t_{n+1}$—then $Y(t)$ will be a semi-Markov process having $\{N(t_n + 0), n = 0, 1, 2, \ldots\}$ for its embedded Markov chain. The sequence of intervals $(t_{n+1} - t_n), n = 0, 1, 2, \ldots$, being the interdeparture time of successive units, forms a renewal process.

Assume that the input process is Poisson with rate λ, and the service times are IID RVs having a general distribution with DF $B(t)$ and mean $(1/\mu)$. Let $B^*(s) = \int_0^\infty e^{-st}\, dB(t)$ be its LST, then $-B^{*(1)}(0) = 1/\mu$.

Let A be the number of arrivals during the service time of a unit. Conditioning on the duration of the service time of a unit, we get

$$k_r = Pr\{A = r\} = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^r}{r!} dB(t), \quad r = 0, 1, 2, \ldots \tag{6.2.1}$$

The PGF $K(s)$ of $\{k_r\}$ is given by

$$\begin{aligned} K(s) &= \sum_{r=0}^{\infty} k_r s^r = \sum_{r=0}^{\infty} s^r \left\{ \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^r}{r!} dB(t) \right\} \\ &= \int_0^\infty e^{-\lambda t}\, dB(t) \left\{ \sum_{r=0}^{\infty} \frac{(\lambda t s)^r}{r!} \right\} \\ &= B^*(\lambda - \lambda s). \end{aligned} \tag{6.2.2a}$$

We have

$$E\{A\} = K'(1) = -\lambda B^{*(1)}(0) = \frac{\lambda}{\mu} = \rho. \tag{6.2.2b}$$

Suppose that the arrivals occur in bulk with distribution $a_j = Pr(X = j)$, the arrival instants being its accordance with a Poisson process with rate λ. Then the probability distribution of the total number of arrivals A in an interval of time t is given by

$$Pr(X = j) = \sum_{k=0}^{j} \frac{e^{-\lambda t}(\lambda t)^k}{k!} a_j^{(k)*}, \quad j = 0, 1, 2, 3, \ldots \tag{6.2.3}$$

where $a_j^{(k)*}$ is the k-fold convolution of a_j with itself. The distribution of the total number of arrivals A during the service period of a unit is given by

$$Pr\{A = j\} = \int_0^\infty \sum_{k=0}^{j} \frac{e^{-\lambda t}(\lambda t)^k}{k!} a_j^{(k)*}\, dB(t). \tag{6.2.4}$$

Let $A(s) = \sum_j a_j s^j$ be the PGF of the bulk size X; then $\sum_j a_j^{(k)*} s^j = [A(s)]^k$.

The PGF of A is then given by

$$\begin{aligned} K(s) &= \sum_{j=0}^{\infty} Pr(A = j) s^j = \sum_{j=0}^{\infty} s^j \left\{ \sum_{k=0}^{j} \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^k}{k!} a_j^{(k)}\, dB(t) \right\} \\ &= \int_0^\infty \sum_{k=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} [A(s)]^k\, dB(t) \\ &= \int_0^\infty e^{-[\lambda t - \lambda A(s) t]}\, dB(t). \end{aligned}$$

Thus, for a bulk arrival system

$$K(s) = B^*[\lambda - \lambda A(s)], \quad \text{and} \tag{6.2.5a}$$

$$\rho = \frac{\lambda E(X)}{\mu}. \tag{6.2.5b}$$

6.3 The M/G/1 Model: Pollaczek-Khinchin Formula

6.3.1 Steady-state distribution of departure epoch system size

Let $t_n, n = 1, 2, \ldots, t_0 = 0$ be the instant at which the nth unit, C_n leaves the system immediately on completion of his service. Then $\{t_n, n \geq 0\}$ is a renewal process. Denote

$$\begin{aligned} N(t) &= \text{number of units at time } t, \\ X_n &= \text{number of units left behind by the } n\text{th departing unit} \\ &= N(t_n + 0), \quad n = 0, 1, 2, \ldots \end{aligned}$$

and

$$A_n = \text{number of units that arrive during the service time of } C_n.$$

We have

$$\begin{aligned} X_{n+1} &= X_n - 1 + A_{n+1}, \quad X_n \geq 1 \\ &= A_{n+1}, \quad X_n = 0. \end{aligned} \tag{6.3.1}$$

Since A_n is the same for all n, we may write $A_n = A$. It can be seen that $\{X_n, n \geq 0\}$ is a Markov chain. It is the embedded Markov chain of the process $\{N(t), t \geq 0\}$ and is the Markov chain extracted from the process at the regeneration points t_n. Denote

$$k_i = Pr\{A = i\}, \quad i = 0, 1, 2, \ldots.$$

Let us find the transition probabilities of the denumerable Markov chain $\{X_n, n \geq 0\}$. Define

$$p_{ij} = Pr\{X_{n+1} = j \mid X_n = i\}. \tag{6.3.2}$$

Then p_{ij} are given by

$$\begin{aligned} p_{ij} &= k_{j-i+1}, \quad i \geq 1, \quad j \geq i - 1 \\ &= 0, \qquad i \geq 1, \quad j < i - 1 \end{aligned}$$

and

$$p_{0j} = p_{ij} = k_j, \quad j \geq 0. \tag{6.3.3}$$

Thus, the TPM of the chain can be put as:

$$\boldsymbol{P} = (p_{ij}) = \begin{bmatrix} k_0 & k_1 & k_2 & \dots \\ k_0 & k_1 & k_2 & \dots \\ 0 & k_0 & k_1 & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}. \tag{6.3.4}$$

This chain has been discussed in Example 1.3, Chapter 1. It can be easily seen that the chain is irreducible and aperiodic. It can also be shown that when $\rho < 1$, the chain is persistent non-null and hence ergodic. We can then apply the ergodic theorem of Markov chains. The limiting probabilities (that a departing customer leaves behind j customers in the system)

$$v_j = \lim_{n\to\infty} p_{ij}^{(n)}, \quad j = 0, 1, 2, \dots$$

exist and are independent of the initial state i. Then the PGF

$$V(s) = \sum_j v_j s^j, \quad \text{of } \{v_j\} \tag{6.3.5}$$

is given by

$$V(s) = \frac{(1 - K'(1))(1 - s)K(s)}{K(s) - s} \tag{6.3.6}$$

(see Example 1.3. Chapter 1).

Using (6.2.2a & 2b) we get, for $0 < \rho < 1$,

$$V(s) = \frac{(1 - \rho)(1 - s)B^*(\lambda - \lambda s)}{B^*(\lambda - \lambda s) - s}. \tag{6.3.7}$$

This is known as the *Pollaczek-Khinchin (P-K) formula.*

Remarks:

(1) What we have obtained is the distribution of the *departure epoch* system size. We shall consider *general time* system size later.

(2) The PGF $Q(s)$ of the number (of customers) in the queue is given by

$$Q(s) = \frac{(1 - \rho)(1 - s)}{B^*(\lambda - \lambda s) - s}.$$

Example 6.1. The $M/D/1$ model
Consider that the service time is Erlang-k (with mean $1/\mu$) having PDF

$$b(t) = \frac{(\mu k)^k t^{k-1} e^{-k\mu t}}{\Gamma(k)}, \quad 0 < t < \infty$$

and LST

$$B^*(s) = \left(\frac{\mu k}{s + \mu k}\right)^k.$$

We get

$$V(s) = \frac{(1-\rho)(1-s)}{1 - s\left\{1 + \frac{\rho(1-s)}{k}\right\}^k}.$$

(See also Section 4.1.1.) The Model $M/D/1$ with constant service time can be considered a limiting case of the above. As $k \to \infty$, the whole mass of the distribution E_k is concentrated at the mean $(1/\mu)$, so E_k can be considered deterministic (constant $= 1/\mu$). Now, as $k \to \infty$, $B^*(s) \to e^{-s/\mu}$ and

$$V(s) \to \frac{(1-\rho)(1-s)}{1 - se^{\rho(1-s)}}.$$

It can be seen that

$$\begin{aligned} v_0 &= 1 - \rho \\ v_1 &= (1-\rho)(e^{\rho} - 1) \\ v_2 &= (1-\rho)(e^{2\rho} - 2(\rho+1)\rho) \end{aligned}$$

and so on.

6.3.2 Waiting-time distribution

Assume that the steady state exists ($\rho < 1$). Let W and W_q denote the waiting times in the system and in the queue, respectively, and let $W(t)$ and $W_q(t)$ denote their distribution functions. We have, for a Poisson input process,

$$p_n = v_n = Pr\{\text{departing unit leaves } n \text{ in the system}\}.$$

Conditioning on the waiting time of the unit, we get

$$\begin{aligned} p_n = \int_0^{\infty} & Pr\{\text{departing unit leaves } n \text{ in the system} \mid t \le W < t + dt\} \\ & \times Pr\{t \le W < t + dt\}. \end{aligned}$$

Again, the event that the departing unit leaves n when his waiting time is t is the event that n arrivals occur during t. Thus,

$$p_n = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} dW(t), \quad n = 0, 1, 2, \ldots. \tag{6.3.8}$$

We have

$$\begin{aligned} V(s) = \sum_{n=0}^{\infty} v_n s^n &= \int_0^\infty \{e^{-\lambda t}\, dW(t)\} \left\{ \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} s^n \right\} \\ &= \int_0^\infty e^{-\lambda t(1-s)}\, dW(t) \\ &= W^*(\lambda - \lambda s), \end{aligned} \tag{6.3.9}$$

where

$$W^*(s) = \int_0^\infty e^{-st}\, dW(t)$$

is the LST of W. Since $p_n = v_n$, we have

$$P(s) = W^*(\lambda - \lambda s).$$

From (6.3.7) and (6.3.9), we get

$$W^*(\lambda - \lambda s) = \frac{(1-\rho)(1-s)B^*(\lambda - \lambda s)}{B^*(\lambda - \lambda s) - s}.$$

Writing α for $\lambda - \lambda s$, we get

$$W^*(\alpha) = \frac{\alpha(1-\rho)B^*(\alpha)}{\alpha - \lambda[1 - B^*(\alpha)]}$$

so that

$$W^*(s) = \frac{s(1-\rho)B^*(s)}{s - \lambda[1 - B^*(s)]}. \tag{6.3.10}$$

This relation connects the LSTs of the distribution of the service time v and the waiting time W in the system. To find the expression for LST $W_q^*(s)$, we note that $W = W_q + v$ (with v being independent of W_q) and

$$W_q^*(s) = \frac{s(1-\rho)}{s - \lambda[1 - B^*(s)]}. \tag{6.3.11a}$$

The formulas (6.3.10) and (6.3.11a) are known as *Pollaczek-Khinchin Transform formulas.*

6.3.2.1 Waiting time in the queue: alternative approach

From renewal theory, we find that the residual service time X of the unit receiving service at the time the test unit arrives has, for large t, the DF

$$G(x) = Pr\{X \leq x\} = \frac{1}{E(v)} \int_0^x \{1 - B(y)\} dy.$$

Its LST equals

$$G^*(s) = \frac{1 - B^*(s)}{s\,E(v)} = \frac{\mu\{1 - B^*(s)\}}{s}.$$

Thus from (6.3.11a) we get

$$W_q^*(s) = \frac{1 - \rho}{1 - \rho G^*(s)}. \tag{6.3.11b}$$

We find that the waiting time, when the service time is general, involves residual service time X, as is to be expected.

Expanding (6.3.11b), we get *Benes' formula*

$$\begin{aligned} W_q^*(s) &= (1 - \rho)\left\{1 + \sum_{k=1}^{\infty} \rho^k [G^*(s)]^k\right\} \\ &= (1 - \rho) + (1 - \rho)\rho G^*(s)/[1 - \rho G^*(s)]. \end{aligned} \tag{6.3.11c}$$

Inversion of (6.3.11b) gives the DF $W_q(x)$ of W_q.

If g denotes the PDF of the residual service times X, then inversion of $[G^*(s)]^k$ gives $g^{(k)*}(x)$, the k-fold convolution of g with itself. Writing $\delta(x)$ as the Dirac delta function and inverting (6.3.11c), we can get the PDF $w_q(x)[=dW_q(x)/dx]$ as follows:

$$w_q(x) = (1 - \rho)\delta(x) + (1 - \rho) \sum_{k=1}^{\infty} \rho^k g^{(k)*}(x), \quad x \geq 0. \tag{6.3.11d}$$

As observed by Kleinrock (1975, p. 201), no satisfactory intuitive explanation has been found for the preceding form of waiting-time PDF in terms of weighted sum of convolved residual service time PDFs.

An interpretation in terms of LIFO-presumptive resume discipline has been advanced in Cooper (1990).

Particular Case. For $M/M/1$, X is again exponential with mean $(1/\mu)$ and

$$g^{k*}(x) = \frac{(\mu^k x^{k-1} e^{-\mu x})}{\Gamma(x)}.$$

Thus, we get

$$w_q(x) = (1-\rho)\delta(x) + (1-\rho)\sum_{k=1}^{\infty} \frac{[(\mu\rho)^k x^{k-1} e^{-\mu x}]}{\Gamma(k)}$$
$$= (1-\rho)\delta(x) + \mu\rho(1-\rho)\exp\{-\mu(1-\rho)x\}. \quad (6.3.11e)$$

Notes:

(1) Shanthikumar (1988) has shown that waiting-time distribution has DFR (decreasing failure rate) for service time having IMRL (increasing mean residual life). He also shows that waiting-time distribution in a $G/G/1$ queue with DFR service time is DFR.

(2) From (6.3.11c) one gets that the LST of the conditional waiting time given that there is a positive wait is given by

$$cW_q^*(s) = (1-\rho)G^*(s)/[1-\rho G^*(s)]$$
$$\text{(since } W_q^*(s) = (1-\rho) + \rho_c W^*(s)).$$

6.3.2.2 Expected waiting time or expected number in the system

The moments $E\{N\}$, $E\{W_q\}$, and $E\{W\}$ can be easily obtained as follows.

We have

$$B^{*(k)}(0) = \frac{d^k}{ds^k} B^*(s)|_{s=0} = (-1)^k E(v^k), \quad k = 1, 2, \ldots.$$

($B^{*(k)}(s)$ denotes the kth derivative of $B^*(s)$). From (6.3.11a), we get

$$\frac{d}{ds}\frac{W_q^*(s)}{1-\rho} = \frac{s - \lambda + \lambda B^*(s) - s(1 + \lambda B^{*(1)}(s))}{[(s - \lambda + \lambda B^*(s)]^2},$$

which takes the form $0/0$ as $s \to 0$. Using L'Hôpital's rule, we get

$$\lim_{s\to o} \frac{d}{ds}\frac{W_q^*(s)}{1-\rho} = \frac{-\lambda B^{*(2)}(0)}{2[1 + \lambda B^{*(1)}(0)]^2} = \frac{-\lambda E(v^2)}{2\left(1 - \frac{\lambda}{\mu}\right)^2},$$

Thus

$$E(W_q) = -W_q^{*(1)}(0) = \frac{\lambda}{2(1-\rho)} E(v^2) \quad (6.3.12a)$$
$$= \frac{\lambda}{2(1-\rho)}\left(\sigma_v^2 + \frac{1}{\mu^2}\right) \quad (6.3.12b)$$
$$= \frac{\rho}{2\mu(1-\rho)}\left(1 + c_v^2\right) \quad (6.3.12c)$$

where σ_v^2 is the variance of the service time v, and c_v is the coefficient of variation of v; that is, $c_v = \mu\sigma_v$. By differentiating k times $W_q^*(s)$ WRT s and putting $s = 0$, we can obtain the kth moment of W_q. It can be seen that

$$\text{var}(W_q) = \frac{\lambda}{12(1-\rho)^2}[4(1-\rho)E(v^3) + 3\lambda\{E(v^2)\}^2]. \quad (6.3.13)$$

The mean waiting time in the system is given by

$$\begin{aligned} E\{W\} &= E\{W_q + v\} = E(v) + E\{W_q\} \\ &= \frac{1}{\mu} + \frac{\lambda}{2(1-\rho)}E(v^2), \qquad (6.3.14) \\ \text{and} \quad \text{var}\{W\} &= \text{var}\{W_q\} + \sigma_v^2. \end{aligned}$$

Using Little's formula, we can obtain $E\{N\}$ from (6.3.14) and $E\{L_q\}$ from (6.3.12). We have

$$\begin{aligned} E\{N\} = \lambda E\{W\} &= \rho + \frac{\lambda^2}{2(1-\rho)}E(v^2) \\ &= \rho + \frac{\rho^2}{1-\rho}\{1 + c_v^2\} \qquad (6.3.15) \\ \text{and} \quad E\{L_q\} = \lambda E\{W_q\} &= \frac{\lambda^2}{2(1-\rho)}E(v^2). \end{aligned}$$

The moments $E\{N\}, E\{N^2\}, \ldots$ can be obtained directly from (6.3.7).

From (6.3.12) and (6.3.14) it follows that for an $M/G/1$ queue (with fixed λ, μ) the expected waiting time (whether in the queue or in the system) increases with the variance of v. It is the least when $\sigma_v = 0$, that is, service-time distribution is constant ($=1/\mu$). The model then becomes $M/D/1$. Thus, determinism in the service time distribution minimizes mean waiting time in a single-server queue with Poisson input. A similar result also holds for a $G/M/1$ queue. This is discussed in Section 6.7.3.

Notes:

(1) The relation (6.3.12a) can also be written as

$$\begin{aligned} E(W_q) &= \frac{\rho}{(1-\rho)} \cdot \frac{E(v^2)}{2E(v)} \\ &= \frac{\rho}{1-\rho}E(v_R), \end{aligned}$$

where v_R is the residual service time of the unit in service.

(2) The first $(k+1)$ moments of the service-time distribution determine the first k moments of the waiting-time distribution and vice versa.

(3) Lemoine (1976) obtains the moments of the waiting time without using the PK formula by utilizing a relationship between the single-server queues and random walks.

(4) For an alternative derivation of the PK formula, see Fakinos (1982).

(5) For higher moments, see Problems and Complements.

(6) *Relations between mean waiting times*

From (6.3.12c) we have, in particular,

$$E(W_q)[M/M/1] = \frac{\rho}{\mu(1-\rho)}$$
$$E(W_q)[M/D/1] = \frac{\rho}{2\mu(1-\rho)}.$$

Thus,

$$E(W_q)[M/G/1] = \{E(W_q)[M/M/1]\}\left(\frac{1+c_v^2}{2}\right) \tag{6.3.12d}$$

and

$$E(W_q)[M/G/1] = c_v^2 E(W_q)[M/M/1] + (1-c_v^2)E(W_q)[M/D/1]. \tag{6.3.12e}$$

Now a question arises whether the above representations of $E(W_q)$ for general service time in terms of that for exponential service time as well as in terms of those of exponential and deterministic (constant) service times would be valid for other more general models. Numerical computations show that such relations do not hold good for general models.

It is found that an approximation

$$E(W_q)[M/G/c] \cong \frac{1}{2}(1+c_v^2)E(W_q)[M/M/c]$$

holds good.

Though (6.3.12d) and (6.3.12e) are not valid for other general models, it is found that such approximations are valid for percentiles of waiting time DF. Define pth percentile $\xi(p)$ of waiting-time DF $W_q(.)$, by

$$W_q(\xi(p)) = p.$$

As $W_q(0) = 1-\rho$, pth percentile $\xi(p)$ is defined for $1-\rho < p < 1$. Corresponding to (6.3.12d) and (6.3.12e), we have the approximation in terms of percentiles as follows:

$$\xi_{gen}(p) = \frac{1}{2}(1+c_v^2)\xi_{\exp}(p),$$
$$\xi_{gen}(p) = c_v^2\xi_{\exp}(p) + (1-c_v^2)\xi_{\det}(p),$$

where $\xi_{gen}(p)$, $\xi_{\exp}(p)$, and $\xi_{det}(p)$ are the pth percentiles for the corresponding service-time distributions.

For a discussion, see Tijms (1994).

Example 6.2. (Sphicas and Shimshak, 1978). Denote

$$c = \frac{SD\{W_q\}}{E\{W_q\}} = \text{coefficient of variation of } W_q \text{ and}$$
$$m_i = E(v^i), \; i\text{th moment of the service time.}$$

We get

$$E(W_q) = \frac{\rho m_2}{2m_1(1-\rho)} \quad \text{and}$$
$$E(W_q^2) = \frac{\rho m_3}{3m_1(1-\rho)} + \frac{\rho^2 m_2^2}{2m_1^2(1-\rho)^2},$$

so that

$$c^2 = \frac{4}{3}\left(\frac{1}{\rho} - 1\right)\left(\frac{m_1 m_3}{m_2^2}\right) + 1 \geq 1.$$

Now $B(.)$ being the DF of service time v, we find that $x\, dB(x)/m_1$ is a PDF with mean m_2/m_1 and variance $[m_1 m_3 - m_2^2]/m_1^2$, so that $m_1 m_3 \geq m_2^2$ and hence,

$$c^2 \geq \frac{4}{3}\left(\frac{1}{\rho} - 1\right) + 1 = \frac{\left(\frac{4}{\rho} - 1\right)}{3} \geq 1.$$

We now examine a particular case: $M/E_k/1$, where

$$m_1 = \frac{1}{\mu}, \quad m_2 = \frac{k+1}{k\mu^2}, \quad m_3 = \frac{(k+1)(k+2)}{k^2\mu^3},$$

so that

$$\frac{\left(\frac{4}{\rho} - 1\right)}{3} \leq c^2 \leq \frac{2}{\rho} - 1.$$

6.3.3 General time system size distribution of an $M/G/1$ queue: supplementary variable technique

We first consider representation of the PDF of a RV through its hazard function.

Suppose that a general distribution G (of a RV v) has the hazard function $r(x)$. Writing $B(x) = Pr\{v \leq x\}$, $R(x) = 1 - B(x)$, we get

$$r(x) = \frac{B'(x)}{1 - B(x)} = \frac{-R'(x)}{R(x)},$$

so that

$$R(x) = \exp\left\{-\int_0^x r(y)dy\right\}$$

and the PDF $f(x)$ of v is given by

$$f(x) = -R'(x) = r(x)\exp\{-N(x)\}$$

where

$$N(x) = \int_0^x r(y)dy$$

$$\left(N(0) = 0 \quad \text{and} \quad \frac{d}{dx}N(x) = r(x)\right).$$

If v is the service time, then $r(x)dx = Pr\{$service will be completed in $(x, x + dx)$ given that service time exceeds $x\}$.

We shall now outline the approach through supplementary variable technique (due to Cox, 1955), which is an important technique for obtaining a transient solution of a non-Markovian system. Inclusion of a supplementary variable enables one to write down the differential equations, as in the case of a Markovian system. The supplementary variable $X(t)$ considered here is defined below.

Let

$N(t)$ = system size at time t

$X(t)$ = time already spent in service by time t of a unit receiving service (spent service time by time t)

$r(x)$ = hazard rate function of the service time

$B(x) = Pr\{v \leq x\}$

$B^*(s)$ = LST of v

$p_n(t) = Pr\{N(t) = n\}$ (with $p_0(0) = 1$)

$p_n(t, x)dx = Pr\{N(t) = n, x \leq X(t) < x + dx\}, \quad n \geq 1$

$$Q(t, z) = \sum_{n=0}^{\infty} p_n(t)z^n$$

$$Q(t, x, z) = \sum_{n=1}^{\infty} p_n(t, x)z^n$$

We have

$$p_n(t) = \int_0^{\infty} p_n(t, x)dx$$

and

$$p_0(t+\delta t) = \{1 - \lambda\delta t + o(\delta t)\}\, p_0(t) + \int_0^\infty p_1(t,x) r(x)\, dx\, \delta t \qquad (6.3.16)$$

(the second term is obtained by conditioning the amount of service already received by the unit in service at time t).

The above equation yields (as $\delta t \to 0$)

$$\frac{\partial}{\partial t}\, p_0(t) = -\lambda p_0(t) + \int_0^\infty p_1(t,x) r(x)\, dx. \qquad (6.3.16a)$$

Again, for $\delta x > 0$

$$p_1(t+\delta t, x+\delta x) = [1 - \lambda\delta t + o(\delta t)][1 - r(x)\delta x + o(\delta x)]\, p_1(t,x). \qquad (6.3.17)$$

Subtracting and adding a term $p_1(t, x+\delta x)$ to the LHS, then dividing by $\delta t(\delta x)$ and taking limits as $\delta t \to 0 (\delta x \to 0)$, we get

$$\frac{\partial}{\partial t}\, p_1(t,x) + \frac{\partial}{\partial x}\, p_1(t,x) = -(\lambda + r(x))\, p_1(t,x). \qquad (6.3.17a)$$

For $n \geq 2$, we shall get an additional term on the RHS of (6.3.17): $p_{n-1}(t,x)$ $(\lambda\delta t)$ [equal to the probability of arrival of a unit in $(t, t+\delta t)$ when there are already $(n-1)$ units in the system]. Thus we shall have, for $n \geq 2$

$$\frac{\partial}{\partial t}\, p_n(t,x) + \frac{\partial}{\partial x}\, p_n(t,x) = -[\lambda + r(x)]\, p_n(t,x) + \lambda p_{n-1}(t,x). \qquad (6.3.18)$$

We shall have the following boundary conditions corresponding to the case where $x = 0+$, with a new service commencing before time t:

$$p_1(t,0) = \int_0^\infty p_2(t,x) r(x)\, dx + \lambda p_0(t) \qquad (6.3.19)$$

and

$$p_n(t,0) = \int_0^\infty p_{n+1}(t,x) r(x)\, dx, \quad n \geq 2. \qquad (6.3.20)$$

Multiplying (6.3.18) by z^n, $n = 2, 3, \ldots$ and (6.3.17a) by z, then adding all the terms, we get

$$\frac{\partial}{\partial t}\left\{\sum_{n=1}^\infty p_n(t,x) z^n\right\} + \frac{\partial}{\partial x}\left\{\sum_{n=1}^\infty p_n(t,x) z^n\right\}$$
$$= -(\lambda + r(x)) \sum_{n=1}^\infty p_n(t,x) z^n + \lambda \sum_{n=2}^\infty p_{n-1}(t,x) z^n,$$

whence

$$\frac{\partial}{\partial t} Q(t, x, z) + \frac{\partial}{\partial x} Q(t, x, z) = -(\lambda - \lambda z + r(x))\, Q(t, x, z). \quad (6.3.21)$$

This is a partial differential equation of the Lagrangian type.

Again multiplying (6.3.20) by z^n, $n = 2, 3, \ldots$ and (6.3.19) by z and adding the terms, one gets

$$Q(t, 0, z) = \int_0^\infty \left\{ \sum_{n=1}^{\infty} p_{n+1}(t, x) z^n \right\} r(x) dx + \lambda z p_0(t). \quad (6.3.22)$$

Now

$$\begin{aligned} &\int_0^\infty \left\{ \sum_{n=1}^{\infty} p_{n+1}(t, x) z^n \right\} r(x) dx \\ &= \left(\frac{1}{z}\right) \int_0^\infty \{Q(t, x, z) - p_1(t, x) z\} r(x) dx \\ &= \left(\frac{1}{z}\right) \left[\int_0^\infty Q(t, x, z) r(x) dx - z\{p_0'(t) + \lambda p_0(t)\} \right] \end{aligned}$$

from (6.3.16a). Thus (6.3.22) reduces to

$$z Q(t, 0, z) = \int_0^\infty Q(t, x, z) r(x) dx - z p_0'(t) + \lambda z(z - 1) p_0(t). \quad (6.3.22a)$$

The partial differential equation (6.3.21) can be solved using the boundary condition (6.3.22a) and the normalizing condition $\sum_{n=0}^{\infty} p_n(t) = 1$.

6.3.3.1 Steady-state distribution of the general time system size

Suppose that

$$\lim_{t \to \infty} p_n(t) = p_n, \quad n \geq 0$$

and

$$\begin{aligned} \lim_{t \to \infty} p_n(t, x) &= p_n(x), \quad x > 0, \ n \geq 1 \\ &= p_0(x) = 0, \quad x > 0 \end{aligned}$$

exist. Then $\{p_n, n \geq 0\}$ gives the distribution of the general time system size.

Let

$$\begin{aligned} Q(x,z) &= \sum_{n=1}^{\infty} p_n(x) z^n \\ &= \sum_{n=1}^{\infty} \{ \lim_{t\to\infty} p_n(t,x) \} z^n \\ &= \lim_{t\to\infty} Q(t,x,z) \end{aligned}$$

and

$$Q(z) = \int_0^{\infty} Q(x,z)\,dx.$$

Then Eqs. (6.3.16a), (6.3.17a)–(6.3.18), (6.3.19), and (6.3.20) reduce, respectively, to

$$\lambda p_0 = \int_0^{\infty} p_1(x) r(x)\,dx \tag{6.3.23}$$

$$\frac{\partial}{\partial x} p_n(x) = -(\lambda + r(x))\, p_n(x) + \lambda p_{n-1}(x), \quad n \geq 1$$

$$(\text{with } p_0(x) = 0, x > 0) \tag{6.3.24}$$

$$p_1(0) = \int_0^{\infty} p_2(x) r(x)\,dx + \lambda p_0 \tag{6.3.25}$$

and

$$p_n(0) = \int_0^{\infty} p_{n+1}(x) r(x)\,dx, \quad n \geq 2. \tag{6.3.26}$$

The partial differential equation (6.3.21) and the boundary condition (6.3.22a) reduce, respectively, to

$$\frac{d}{dx} Q(x,z) = -[\lambda - \lambda z + r(x)]\, Q(x,z) \tag{6.3.27}$$

and

$$z Q(0,z) = \int_0^{\infty} Q(x,z) r(x)\,dx + \lambda z(z-1)\, p_0; \tag{6.3.28}$$

the normalizing condition is

$$p_0 + Q(1) = 1. \tag{6.3.29}$$

Equation (6.3.27) is a linear first-order differential equation whose solution is

$$Q(x,z) = Q(0,z) \exp\{-\lambda(1-z)x - N(x)\}. \tag{6.3.30}$$

Substituting the above expression for $Q(x, z)$ in (6.3.28), we get

$$\begin{aligned} zQ(0,z) &= Q(0,z)\int_0^\infty \exp\{-\lambda(1-z)x - N(x)\}r(x)dx + \lambda z(z-1)p_0 \\ &= Q(0,z)\left[\int_0^\infty e^{-\lambda(1-z)x}\left\{e^{-N(x)}r(x)\right\}dx\right] + \lambda z(z-1)p_0 \\ &= Q(0,z)B^*(\lambda-\lambda z) + \lambda z(z-1)p_0, \end{aligned}$$

so that

$$Q(0,z) = \frac{\lambda z(z-1)p_0}{z - B^*(\lambda - \lambda z)}. \tag{6.3.31}$$

We then get from (6.3.30)

$$\begin{aligned} Q(z) &= \int_0^\infty Q(x,z)dx \\ &= Q(0,z)\int_0^\infty \exp\{-\lambda(1-z)x - N(x)\}dx \\ &= Q(0,z)\int_0^\infty e^{-N(x)}e^{-\lambda(1-z)x}\,dx. \end{aligned}$$

Integrating by parts, we get

$$\begin{aligned} Q(z) &= \frac{Q(0,z))}{\lambda(1-z)}\left[1 - \int_0^\infty e^{-\lambda(1-z)x}\left\{e^{-N(x)}r(x)\right\}dx\right] \\ &= \frac{Q(0,z)}{\lambda(1-z)}[1 - B^*(\lambda-\lambda z)]. \end{aligned} \tag{6.3.32}$$

Thus, from (6.3.31) and (6.3.32), we get

$$Q(z) = \frac{z[B^*(\lambda-\lambda z) - 1]p_0}{z - B^*(\lambda-\lambda z)}. \tag{6.3.33}$$

Using L'Hôpital's rule, we get

$$\begin{aligned} Q(1) &= \lim_{z\to 1} Q(z) \\ &= p_0 \lim_{z\to 1} \frac{[B^*(\lambda-\lambda z) - 1] + z[-\lambda B^{*(1)}(\lambda-\lambda z)]}{1 + \lambda B^{*(1)}(\lambda-\lambda z)} \\ &= p_0 \frac{\rho}{1-\rho}, \quad \text{since } -B^{*(1)}(0) = E(v) = 1/\mu. \end{aligned}$$

The normalizing condition

$$p_0 + Q(1) = 1$$

yields

$$p_0 = 1 - \rho$$

(a result that holds for any single-server queueing system in steady state). Finally, we get from (6.3.33)

$$Q(z) = \frac{(1-\rho)z[B^*(\lambda - \lambda z) - 1]}{z - B^*(\lambda - \lambda z)}. \tag{6.3.34}$$

Now the PGF of the general time system size in steady state is given by

$$P(z) = \sum_{n=0}^{\infty} p_n z^n = p_0 + \sum_{n=1}^{\infty} p_n z^n = p_0 + Q(z)$$
$$= (1-\rho)\frac{B^*(\lambda - \lambda z)(1-z)}{B^*(\lambda - \lambda z) - z}.$$

Thus, $P(z) = V(z)$ where

$$V(z) = \sum_{n=0}^{\infty} v_n z^n$$

is the Pollaczek-Khinchin formula (for departure-epoch system-size distribution). We find that in steady state the general time system size has the same distribution as the departure epoch system size.

Remark: Incidentally, we have established the equality $P(z) = V(z)$—that is, $v_n = p_n$ for all $n \geq 0$. This also follows from the result that in a system with Poisson input PASTA holds and so $v_n = p_n$ (also noted in (1) Section 6.2). Further, d_n and v_n denote the same limiting probability.

The equality does not hold if the input is from a finite source or the system is a finite buffer system (with limited waiting space).

Notes:

(1) See Cox (1955), Keilson and Kooharian (1960), and Cohen (1982) for details of the supplementary variable technique applied to $M/G/1$ queue. Borthakur and Medhi (1974) apply the technique to a more general model $M^X/G(a,b)/1$.

(2) Characterization problems of equilibrium system size distribution in $M/G/1$ queues have been studied. Rego (1988) shows that of all $M/G/1$ queues, the $M/M/1$ queue is the only one in which both $\{p_j\}$ and $\{k_j\}$ have geometric distributions. Disney *et al.* (1973) studied characterization through renewal departure processes; they showed that the departure process from an $M/G/1$ queue with infinite capacity is a renewal process *iff* it is in

steady state and service times are exponential (i.e., the queue is an $M/M/1$ queue).

6.3.4 Semi-Markov process approach

The system-size $\{N(t)\}$ is not semi-Markovian. Consider that a transition occurs with a service completion (departure of a unit)— that is, t_n, $n = 0, 1, 2, \ldots$ are the nth departure epochs. Using the notation of Section 1.9,

$$\begin{aligned} X_n &= N(t_n + 0) = \text{system size at the } n\text{th departure} \\ &\quad\ (\text{i.e., number left behind by the } n\text{th departure}) \\ Y(t) &= X_n, \quad t_n \le t < t_{n+1}, \end{aligned}$$

we see that $\{Y(t), t \ge 0\}$ is a semi-Markov process that has $\{X_n, n \ge 0\}$ for its embedded Markov chain. We get

$$\begin{aligned} Q_{i,j}(t) &= 0, \quad i \ge 1, \ j < 1, \ j < i - 1 \\ &= \int_0^t \frac{e^{-\lambda y}(\lambda y)^{j-i+1}}{(j-i+1)!} dB(y), \quad i \ge 1, \ j \ge i - 1 \\ &= \int_0^t \left\{1 - e^{-\lambda(t-y)}\right\} \frac{e^{-\lambda y}(\lambda y)^j}{j!} dB(y), \quad i = 0, \ j \ge 0. \end{aligned}$$

One can proceed, as in Section 1.9, to find

$$v_j = \lim_{n\to\infty} p_{ij}^{(n)}$$

and

$$p_j = \lim_{t\to\infty} Pr\{Y(t) = j\}.$$

See Fabens (1961) for a relationship between v_j and π_j, where

$$\pi_j = \lim_{t\to\infty} Pr\{N(t) = j\}.$$

See also Neuts (1967) for semi-Markov analysis of the more general model $M/G(a, b)/1$.

6.3.5 Approach via martingale

Approach via martingale is also being considered in the literature; such an approach leads to interesting results. Baccelli and Makowski (1989) consider an exponential martingale associated with the (embedded) Markov chain $\{X_n, n \ge 0\}$, where X_n is the number left behind by the nth departure in an $M/G/1$ queue in steady state. Using the basic regularity properties of martingales

(Baccelli and Makowski, 1986) and standard renewal theoretic arguments, they show that this martingale provides a unified probabilistic framework for deriving several interesting well-known results, such as the Pollaczek-Khinchin formula, the transient generating function of the departure epoch system size, and the generating function of the number served during a busy period. Rosenkrantz (1983) obtains the busy period distribution of an $M/G/1$ queue via a martingale approach. (See Remark (5), Section 6.4.2.)

Example 6.3. An $M/G/1$ queue with a second optional channel
It may happen that as soon as the first essential service of a unit is complete, it may leave the system with probability $(1-r)$ or may immediately opt for a second optional service with probability $r(0 \leq r \leq 1)$. Consider that the service times v_1, v_2 of the two channels are independent, having general distributions, with DF $B_i(\cdot)$, LST $B_i^*(\cdot)$ and mean service times $b_i, i = 1, 2$ (denoting the first and second channel, respectively). Assume that the same server serves both the channels, or if there are two different servers, only one unit can be in service in either channel at any time and that no other unit can be taken in for service until the unit being served leaves the system after final service completion. Then the service time of a unit is

$$\begin{aligned} v &= v_1 + v_2, \text{ with probability } r \\ &= v_1, \text{ with probability } 1 - r. \end{aligned}$$

The LST $B^*(s)$ of v is

$$B^*(s) = (1-r)B_1^*(s) + r\,B_1^*(s)\,B_2^*(s).$$

The moments are

$$\begin{aligned} b \equiv E(v) &= b_1 + r b_2 \\ b^{(2)} \equiv E(v^2) &= b_1^{(2)} + 2r b_1 b_2 + r b_2^{(2)}, \\ \text{where} \quad b_i^{(2)} &= E(v_i^2), \quad i = 1, 2, \quad \text{and} \\ \rho &= \lambda b = \lambda(b_1 + r b_2). \end{aligned}$$

One can thus get the corresponding results for this model by writing the expression of $B^*(s)$ given above, including the transient solution, the busy period distribution, and when $\rho < 1$, one can get the steady state Pollaczek-Khinchin formula (see Medhi (2002)).

Consider the particular case when the service time of the second channel is exponential (case considered in Madan (2000)). Then

$$B^*(s) = B_1^*(s)\left[\frac{1 + (1-r)b_2 s}{1 + b_2 s}\right]$$

and the Pollaczek-Khinchin formula becomes

$$\begin{aligned} V(s) &= \sum_{i=0}^{\infty} v_i s^i \\ &= \frac{(1-\rho)(1-s)B_1^*(\lambda-\lambda s)}{B_1^*(\lambda-\lambda s)} \\ &\quad \cdot \frac{[1+(1-r)(\lambda-\lambda s)b_2]}{[1-(1-r)(\lambda-\lambda s)b_2 - s\{1+(\lambda-\lambda s)b_2\}}. \end{aligned}$$

The mean number $E(N)$ in the system is given by

$$E\{N\} = \rho + \frac{\lambda^2\left[b_1^{(2)} + 2r(b_1+b_2)b_2\right]}{2(1-\rho)b_2^2}$$

(Medhi, 2002).

6.4 Busy Period

6.4.1 Introduction

Assume that a busy period is initiated by a single customer. It commences at the instant of the arrival of a customer to an empty system and terminates at the instant the server becomes free for the first time. The initiating customer will be called the "ancestor," and the customers who arrive during the service time of the initiating customer will be called its "descendants." The ith customer who arrives during the service time of the initiating customer will be called the ith descendant. The descendants of the ith descendant will be called the "progenies" of the ith descendant. As usual, let

$$\begin{aligned} T &= \text{duration of the busy period} \\ G(t) &= Pr\{T \le t\} \\ B(t) &= Pr\{v \le t\}, \quad \text{where } v \text{ is the service time} \\ G^*(s) &= \text{LST of } T(=E\{e^{-sT}\}) \\ B^*(s) &= \text{LST of } v(=E\{e^{-sv}\}) \\ G^{(n)*}(t) &= n\text{-fold consolution of } G(t) \text{ with itself} \\ A(t) &= \text{number of arrivals during}(0, t) \\ &\quad (\equiv \text{a Poisson random variable with mean } \lambda t) \\ g_r &= E\{T^r\} \\ T_i &= \text{total service time of the } i\text{th descendant and all its progeny} \end{aligned}$$

It can be seen that T_i is a busy period initiated by a single customer (the ith customer), and as such T_i has the same distribution as the busy period T. T_i is called a *sub-busy* (or *pseudo-busy*) period. The T_i's are IID random variables distributed as T.

6.4.2 Busy-period distribution: Takács integral equation

The distribution of a busy period was obtained by Takács (1962). The result is discussed below.

Theorem 6.1. *The LST of the busy period can be expressed as a functional equation*

$$G^*(s) = B^*[s + \lambda - \lambda G^*(s)]. \tag{6.4.1}$$

Proof: To obtain the busy period distribution, we condition on two events—namely, on the duration of the service time v of the initiating customer (ancestor) and on the number A of arrivals during the service time of the ancestor. Given that $v = x$ and $A \equiv A(x) = n$, then n sub-busy periods $T_1, \ldots, T_n$ are generated by the n descendants, and

$$T = x + T_1 + \cdots + T_n. \tag{6.4.2}$$

Since the T_i's are IID as T and are also independent of x, we have

$$\begin{aligned} E\{e^{-sT} \mid v = x, A = n\} &= E\{e^{-s(x+T_1+\cdots+T_n)}\} \\ &= [E\{e^{-sx}\}]\left[E\{e^{-s(T_1+\cdots+T_n)}\}\right] \\ &= e^{-sx}[G^*(s)]^n \end{aligned} \tag{6.4.3}$$

(since x is a fixed quantity). To get the LST of T, we have to remove the conditions on v and A. We have

$$\begin{aligned} E\{e^{-sT} \mid v = x\} &= \sum_{n=0}^{\infty} E\{e^{-sT} \mid v = x, A = n\} Pr\{A = n\} \\ &= \sum_{n=0}^{\infty} e^{-sx}[G^*(s)]^n \frac{e^{-\lambda x}(\lambda x)^n}{n!} \\ &= e^{-(s+\lambda-\lambda G^*(s))x}. \end{aligned} \tag{6.4.4}$$

Finally,

$$\begin{aligned} E\{e^{-sT}\} &= \int_0^{\infty} E\{e^{-sT} \mid v = x\} dB(x) \\ &= \int_0^{\infty} e^{-(s+\lambda-\lambda G^*(s))x} dB(x). \end{aligned}$$

Thus,

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)). \qquad \blacksquare$$

Particular Case. For an $M/M/1$ system, $B^*(s) = \mu/(s+\mu)$ and then (6.4.1) reduces to a quadratic in $G^*(s)$; its unique root (such that $G^*(0) = 1$) is z_2/ρ. (See Eqs. (3.9.5) and (3.9.26), Chapter 3.)

Remarks:

(1) It was Good who first noted the possibility of analyzing a busy period as a branching process.

(2) The functional equation has a unique root $G^*(s) \leq 1$. Further, $G^*(s)$ is the LST of a proper probability distribution *iff* $\rho \leq 1$. (For a proof, see Example, III (4.a), p. 417, Feller, Vol. II (1966).)

(3) *Takács integral equation.* We have

$$\begin{aligned} &Pr\{T \leq t \mid v = x,\ A = n\} \\ &\quad = Pr\{T_1 \cdots + T_n \leq t - x\} \\ &\quad = G^{(n)*}(t - x), \end{aligned}$$

since T_i's are independent. Thus,

$$G(t) = Pr\{T \leq t\} = \int_0^\infty \sum_{n=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^n G^{(n)*}(t-x)}{n!} dB(x), \qquad (6.4.5)$$

which is known as Takács integral equation for the busy period. Taking LST of the above and interchanging the order of integration, on simplification one gets the functional equation (6.4.1).

(4) Conway *et al.* (1967) show how one can use the analysis of a busy period to obtain the distribution of the waiting time in the system.

(5) Rosenkrantz (1983) gives a new formula for $G^*(s)$ through a martingale approach. The formula that expresses $G^*(s)$ in terms of $B^*(s)$ is of independent interest in itself. Writing

$$f(s) = \lambda[B^*(s) - 1] + s$$

and

$$f(f^{-1}(s)) = f^{-1}(f(s)) = s$$

($f^{-1}(s)$, the inverse of $f(s)$ is assumed to exist in the neighborhood of 0), he shows that

$$G^*(f(s)) = B^*(s); \qquad (6.4.6)$$

or, taking inverses

$$G^*(s) = B^*(f^{-1}(s)). \tag{6.4.7}$$

The formula immediately yields $E(T)$ and higher moments of T.

(6) The expected busy period $E(T)$ can be obtained even without finding the distribution of T by applying a result from renewal theory (see Remark 4, Section 3.9.4). The same result holds since the input is Poisson (and thus $E(I) = 1/\lambda$).

Consider an $M/G/1$ queue.

Define a cycle C as the time elapsed between two consecutive arrivals who find the system empty. Then

$$C = I + T.$$

Denote

$C_j \equiv$ amount of time that j customers are present during one cycle,

$p_j =$ long-run fraction of time that j customers are present.

$=$ Prob {the number in the system $= j$}, $j = 0, 1, 2, \ldots$.

Then

$$p_j = \frac{E(C_j)}{E(C)}, \quad j = 0, 1, 2, \ldots.$$

In particular, $E(C_0) = E(I) = 1/\lambda$ in case of Poisson arrivals.

Thus,

$$p_0 = \frac{1}{\lambda E(C)} = \frac{1}{\lambda[E(I) + E(T)]}.$$

But $p_0 = 1 - \rho$, whence

$$E(T) = \frac{1}{\mu(1-\rho)}.$$

However, this is not the case when the input is not Poisson.

6.4.3 Further discussion of the busy period

6.4.3.1 Moments of the busy period

The functional equation (6.4.1) is difficult to invert. However, one can easily find the moments of T. We have

$$\begin{aligned} E(T) &= -\frac{d}{ds} G^*(s)|_{s=0} \\ &= E(v)\{1 + \lambda E(T)\}, \end{aligned}$$

so that

$$E(T) = \frac{E(v)}{(1-\rho)} = \frac{1}{\mu(1-\rho)} = \frac{\rho}{1-\rho}\{E(I)\}. \tag{6.4.8}$$

The expected duration of the busy period of an $M/G/1$ queue is the same as that of an $M/M/1$ queue. It is independent of the form of the service-time distribution (that occurs only through its mean).

$$\begin{aligned} E(T^2) &= \frac{d^2}{ds^2} G^*(s)|_{s=0} \\ &= B^{*(2)}(0)\left[1 - \lambda G^{*(1)}(0)\right]^2 + B^{*(1)}(0)\left[-\lambda G^{*(2)}(0)\right] \\ &= E(v^2)[1 + \lambda E(T)]^2 + \lambda E(v) E(T^2). \end{aligned}$$

Thus,

$$\begin{aligned} E(T^2) &= \frac{E(v^2)[1 + \lambda E(T)]^2}{1 - \lambda E(v)} \\ &= \frac{E(v^2)}{(1-\rho)^3}. \end{aligned} \tag{6.4.9}$$

We have

$$\text{var}(T) = \frac{E(v^2)}{(1-\rho)^3} - \left[\frac{1}{\mu(1-\rho)}\right]^2 = \frac{1}{(1-\rho)^3}\left[\sigma_v^2 + \rho(E(v))^2\right]. \tag{6.4.10}$$

6.4.3.2 Number served during a busy period

$N(T) \equiv N =$ number served during a busy period (T)

$N_i =$ number served during the sub-busy period initiated by the ith descendant

$=$ number served during the sub-busy period T_i

(i.e., total service time of the ith descendant and all its progeny)

$$P(z) = \sum_{k=1}^{\infty} Pr(N = k) z^k \text{ (PGF of } N \text{ or } N(T))$$

$$h_r = E(N^r), \quad r = 1, 2, \ldots$$

It is clear that all the N_i's are independently distributed as N, and

$$P_i(z) = \sum_{k=1}^{\infty} Pr(N_i = k) z^k = P(z)$$

for all i. Given that the number of arrivals A during the service time of the ancestor is n, we have

$$N = 1 + N_1 + \cdots + N_n. \tag{6.4.11}$$

Because of the independence of the N_i's, the conditional generating function of N can be written as

$$\begin{aligned} E(z^N \mid A = n) &= E\left(z^{1+N_1+\cdots+N_n}\right) \\ &= E(z)\prod_{i=1}^{n} E\left(z^{N_i}\right) \\ &= z[P(z)]^n. \end{aligned} \tag{6.4.12}$$

Removing the condition on A, we get

$$\begin{aligned} P(z) = E(z^N) &= \sum_{n=0}^{\infty} E(z^N \mid A = n) Pr\{A = n\} \\ &= z\sum_{n=0}^{\infty} [P(z)]^n Pr\{A = n\}. \end{aligned} \tag{6.4.13}$$

We have

$$\begin{aligned} &\sum_{n=0}^{\infty} Pr(A = n) z^n \\ &\quad = \text{PGF of the number of arrivals during the service time of a unit} \\ &\quad = B^*(\lambda - \lambda z) \text{ for a Poisson input system,} \end{aligned} \tag{6.4.14}$$

so that

$$\sum_{n=0}^{\infty} Pr\{A = n\}[P(z)]^n = B^*(\lambda - \lambda P(z)). \tag{6.4.15}$$

Finally, we get from (6.4.13) and (6.4.15)

$$P(z) = zB^*(\lambda - \lambda P(z)). \tag{6.4.16}$$

The PGF $P(z)$ of the number served during a busy period satisfies the above functional equation.

Notes:

(1) The joint bivariate distribution of T and N of an $M/G/1$ system has been obtained by Prabhu (1960, 1965). Enns (1969) and Scott and Ulmer (1972) consider a joint trivariate distribution of T, N, and M (the maximum number served during a busy period).

(2) Busy period of an $M/G/1/K$ queue has been considered by Harris (1971) and Miller (1975) (see Problems and Complements 6.10).

(3) Shanthikumar (1988) shows that the number served during a busy period is DFR for IFR service times.

(4) The number N served during a busy period can also be expressed as a first passage time of a certain Markov chain with an absorbing state 0.

(5) Let $G^*(z, s)$ be the two-dimensional transform of the number served during a busy period, and the duration of the busy period.

For $z = 1$, $G^*(1, s) = G^*(s)$ is the LST of the busy period, and for $s = 0$, $G^*(z, 0) = P^*(z)$ is the PGF of the number served during a busy period.

$G^*(z, s)$ satisfies the relation

$$G^*(z, s) = zB^*(s + \lambda - \lambda G^*(z, s)).$$

6.4.3.3 Moments of the number served during a busy period

The functional equation (6.4.16) is difficult to solve. However, moments of N can be obtained easily from it. Note that

$$\begin{aligned}\frac{d^k}{dz^k} B^*(z)|_{z=0} &= B^{*(k)}(0) \\ &= (-1)^k \mu_k,\end{aligned}$$

where $\mu_k = E(v^k) = k$th moment of the service time distribution (with $\mu_1 = 1/\mu$). We get from (6.4.16)

$$\begin{aligned}h_1 &= P'(1) \\ &= B^*(0) - \lambda B^{*(1)}(0) P'(1) \\ &= B^*(0) + \frac{\lambda}{\mu} h_1 = 1 + \rho h_1,\end{aligned}$$

whence

$$h_1 = \frac{1}{1 - \rho}. \tag{6.4.17}$$

Further,

$$\begin{aligned}P''(1) &= B^{*(1)}(0)[-\lambda P'(1)] + B^{*(1)}(0)[-\lambda P'(1)] + B^{*(2)}(0)[-\lambda P'(1)]^2 \\ &\quad + B^{*(1)}(0)[-\lambda P''(1)] \\ &= 2\frac{\lambda}{\mu} h_1 + \mu_2(-\lambda h_1)^2 + \frac{\lambda}{\mu} P''(1),\end{aligned}$$

whence

$$P''(1) = \frac{2\rho(1-\rho) + \lambda^2\mu_2}{(1-\rho)^3}.$$

Since $P''(1) = h_2 - h_1$, we get

$$h_2 = \frac{2\rho(1-\rho) + \lambda^2\mu_2}{(1-\rho)^3} + \frac{1}{1-\rho}. \tag{6.4.18}$$

Thus, the variance of the number served during a busy period is given by

$$\sigma_h^2 = \frac{\rho(1-\rho) + \lambda^2\mu_2}{(1-\rho)^3}. \tag{6.4.19}$$

Particular Case. (a) *The model* $M/M/1$. Here

$$\mu_1 = \frac{1}{\mu}, \qquad \mu_2 = \frac{2}{\mu^2}, \qquad B^*(s) = \frac{\mu}{s+\mu}.$$

Then the functional equation (6.4.16) reduces to

$$P(z) = \frac{z\mu}{\mu + \lambda - \lambda P(z)}, \quad \text{that is,}$$

$$\lambda P^2(z) - (\lambda + \mu) P(z) + \mu z = 0. \tag{6.4.20}$$

Solving and considering the root for which $P(1) = 1$, we get,

$$P(z) = \frac{1+\rho}{2\rho}\left\{1 - \left[1 - \frac{4\rho z}{(1+\rho)^2}\right]^{1/2}\right\}. \tag{6.4.21}$$

Expanding $P(z)$ in a power series in z and comparing the coefficients of z^n, we get

$$Pr[N(T) = n] = \frac{\frac{1}{n}\binom{2n-2}{n-1}\rho^{n-1}}{(1+\rho)^{2n-1}}, \quad n \geq 1 \tag{6.4.22}$$

is the distribution of the number $N(T)$ served during a busy period T.

This distribution has been described by Haight (1961) as "analogous to the Borel-Tanner." The mean and variance of the number served during a busy period can be obtained from (6.4.17) and (6.4.19) or directly from (6.4.21). These are given by

$$E\{N(T)\} = \frac{1}{1-\rho} \quad \text{and} \tag{6.4.23}$$

$$\text{var}\{N(T)\} = \frac{\rho(1+\rho)}{(1-\rho)^3}. \tag{6.4.24}$$

(b) *The Model M/D/1.* We have $\mu_1 = 1/\mu, \mu_2 = 0$, $B^*(s) = e^{-s/\mu}$ so that the functional equation becomes

$$\begin{aligned} P(z) &= ze^{-[\lambda-\lambda P(z)]/\mu} \\ &= ze^{-\rho}e^{\rho P(z)}. \end{aligned} \tag{6.4.25}$$

This is a particular case of what is known as a Borel-Tanner distribution. It can be shown that

$$Pr[N(T) = n] = \frac{(n\rho)^{n-1}e^{-n\rho}}{n!}, \quad n = 1, 2, \ldots. \tag{6.4.26}$$

The mean and variance can be easily obtained from (6.4.25). These are given by

$$E\{N(T)\} = \frac{1}{(1-\rho)} \tag{6.4.27}$$

$$\text{var}\{N(T)\} = \frac{\rho}{(1-\rho)^2}. \tag{6.4.28}$$

In this particular case, we can easily obtain the DF $G(t)$ of the busy period by using (6.4.26). For, if the number served during the busy period T is n, then the duration of the busy period is n/μ, since the service time of each of the customers is of constant duration $1/\mu$. Thus,

$$G(t) = Pr\{T \le t\} = Pr\left\{\frac{n}{\mu} \le t\right\} = Pr\{n \le \mu t\}$$

$$= \sum_{n=1}^{(\mu t)} \frac{(n\rho)^{n-1}e^{-n\rho}}{n!}$$

(using (6.4.26)), where $[\mu t]$ is the greatest integer not exceeding μt.

6.4.4 Delay busy period

Often it is useful to consider a more general kind of busy period that is initiated by the perfomance of some initial task before starting service (Conway *et al.*, 1967). The time T_0 is the time spent in the initial task and the time needed for service of the first unit. The time T_b spent in servicing units (subsequent to T_0) until none is left is the ordinary busy period. When there is no initial task, the ordinary busy period is denoted by T. Let $G_0^*(s)$, $G_b^*(s)$, $G_c^*(s)$ be the LST of T_0, T_b and $T_c(=T_0 + T_b)$, respectively, and $G^*(s)$ be the LST of ordinary busy period T (with no initial task).

Consider that the expected number of arrivals during T_0 is $\lambda E(T_0)$ and that the delay busy period is initiated by the number of arrivals during T_0. It can be

seen that

$$E(T_b) = \lambda E(T_0)[E(T)]$$
$$= \frac{\rho}{1-\rho} E(T_0). \tag{6.4.29}$$

Further, if we denote the duration of the busy cycle by T_c, so that

$$T_c = T_0 + T_b,$$

then

$$E(T_c) = E(T_0) + E(T_b)$$
$$= \frac{E(T_0)}{1-\rho}. \tag{6.4.30}$$

It is clear that $T_b = \sum_{i=1}^{N} T$, where N is the RV denoting the number of arrivals during T_0 and is independent of T. Now N has the PGF $G_0^*(\lambda - \lambda s)$ since arrivals are Poisson. T_b is the sum of a random number of RVs each equal to T. Hence, the LST of T_b is given by

$$G_b^*(s) = G_0^*[\lambda - \lambda G^*(s)]. \tag{6.4.31}$$

Further, since the first task (the initiating task) of the delay busy period has LST $G_0^*(s)$, we can at once obtain the LST of T_c in the same manner as that obtained for T. We can thus write down the LST of T_c by writing G_0 in place of B on the RHS expression for $G^*(s)$ in (6.4.1). Thus,

$$G_c^*(s) = G_0^*[s + \lambda - \lambda G^*(s)]. \tag{6.4.32}$$

The mean is given by

$$E(T_c) = -\frac{d}{ds}[G_c^*(s)]|_{s=0};$$

on simplification, one gets

$$E(T_c) = \frac{E(T_0)}{1-\rho},$$

as is given by (6.4.30).

6.4.5 Delay busy period under *N*-policy

It is generally assumed that when the server completes his service on a customer and finds no one waiting to serve, he remains idle until the next customer arrives. However, instead of the server commencing service as soon as one customer arrives, the server may wait until the queue length reaches a desired level $N(\geq 1)$. In other words, each time the system becomes empty, the server waits (or gets busy with other work) until the queue length becomes exactly N,

then commences serving customers one by one and continues until the system becomes completely empty. This is called *N-policy* with *exhaustive service.* The period the server waits until the arrival of the Nth customer is also called a *server's vacation.* (See Section 8.3.) The length of a vacation depends on the arrival process. The model was studied by Yadin and Naor (1963) and Heyman (1968). The latter showed that this model possesses certain optimal properties. (See Section 8.4.)

It may happen that the server returning from vacation when the queue length builds up to exactly N may not be immediatly available for servicing the waiting customers. He may be engaged in some preservice work or in gearing up the service mechanism for service operation. (This kind of situation may also arise in an inventory problem.) The time that the server remains occupied in such preservice work may be called start-up time (SUT). This model has been discussed earlier (see Problem 3.3 in Chapter 3) and by Borthakur *et al.* (1987) as a more generalized model. By considering that SUTs are identically zero, one may get the models earlier studied. While Baker analyzes the steady-state behavior of an $M/M/1$ queue under N-policy and exponential SUT, Borthakur *et al.* consider an $M/M/1$ queue with N-policy and general SUT.

We consider here an extension of the Takács integral equation for busy period of an $M/G/1$ model with N-policy and general SUT (Medhi and Templeton, 1992). Here the busy period is defined as the length of the interval B from the instant that the Nth customer arrives in an (server) idle system to the instant the server completes servicing of all the units, leaving the system empty for the first time. The RV B is the sum of two RVs: (a) the SUT denoted by U and (b) the period T during which the server is occupied with actual servicing of units consisting of (i) N units, (ii) those that arrive during the start-up time, and (iii) those that arrive during the service times of customers (i) and (ii).

Thus, $B = U + T$, where U and T are not independent. Hence, B is the busy period of an $M/G/1$ queue initiated by N customers, plus those customers who arrive during the SUT U.

Let λ and μ be the arrival and service rates $G_0^*(s) =$ LST of U, with $E(U) = u$, $G^*(s) =$ LST of a busy period V initiated by one customer in a standard $M/G/1$ queue, with DF $G(x)$, and $G_B^*(s) =$ LST of the busy period B with DF $G_B(x)$. By conditioning on the duration of the SUT, we get,

$$G_B(x) = Pr(B \leq x) = \int_0^x Pr\{B_1 \mid U = t\} dU(t),$$

where $(B_1 \mid U = t)$ is the event that the busy period generated by N customers plus those arriving during SUT of length t is less than or equal to $(x - t)$. Thus,

$$G_B(x) = \int_0^x \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} G^{(N+n)*}(x - t) dU(t), \qquad (6.4.33)$$

where $G^{(N+n)*}$ is the $(N + n)$-fold convolution of G with itself.

Taking the LST of (6.4.33), one gets

$$G_B^*(s) = \int_0^\infty \int_0^x \sum_{n=0}^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} e^{-sx} G^{(N+n)*}(x-t)dU(t)dx. \quad (6.4.34)$$

Changing the order of integration, one gets

$$\begin{aligned} G_B^*(s) &= \int_0^\infty \sum_{n=0}^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} dU(t) \times \int_t^\infty e^{-sx} G^{(N+n)*}(x-t)dx \\ &= \int_0^\infty \sum_{n=0}^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!} e^{-st}[G^*(s)]^{N+n} dU(t) \\ &= [G^*(s)]^N \int_0^\infty e^{-\lambda t} e^{\lambda t G^*(s)} e^{-st} dU(t) \\ &= [G^*(s)]^N \int_0^\infty e^{-[s+\lambda-\lambda G^*(s)]t} dU(t). \end{aligned}$$

Thus,

$$G_B^*(s) = [G^*(s)]^N G_0^*[s + \lambda - \lambda G^*(s)], \quad N \geq 1, \quad (6.4.35)$$

where $G^*(s)$ is given by the Takács integral equation

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)),$$

B^* being the LST of the service-time distribution. We can at once obtain the moments of B from the preceding. From (6.4.35), we get

$$\begin{aligned} E(B) &= -\frac{d}{ds} G_B^*(S)|_{s=0} = NE(V) + E(U)[1 + \lambda E(V)] \\ &= \frac{N}{\mu(1-\rho)} + \frac{u}{1-\rho}, \end{aligned} \quad (6.4.36)$$

since

$$E(V) = \frac{1}{\mu(1-\rho)}, \quad \rho = \frac{\lambda}{\mu}.$$

Remarks:

(1) By considering that, on an average, the number of arrivals during U is λu, it can be seen that

$$E(T) = (N + \lambda u)E(V),$$

so that

$$\begin{aligned} E(B) &= E(U) + E(T) \\ &= u + \frac{N + \lambda u}{\mu(1-\rho)} \\ &= \frac{N}{\mu(1-\rho)} + \frac{u}{1-\rho}. \end{aligned}$$

(2) If I is the idle period, then $E(I) = N/\lambda$. Denote by p_0 the steady-state probability that the server is idle with a number in the queue less than N.

Then using

$$p_0 = \frac{E(I)}{E(I) + E(B)},$$

we get

$$p_0 = \frac{N(1-\rho)}{N + \lambda u}. \tag{6.4.37}$$

as the fraction of time the server remains idle. The fraction of time the server remains busy (either with preservice work or with servicing of units) equals

$$1 - p_0 = \rho + \frac{\lambda u(1-\rho)}{N + \lambda u}.$$

Another interesting point is that when $u = 0$ (i.e., start-up time is zero), then $1 - p_0 = \rho$, independent of the control parameter N.

(3) Equation (6.4.32) gives the LST of the delay busy period of the system where the server becomes engaged in start-up work as soon as the system becomes empty. Equation (6.4.35) gives the LST of the delay busy period of the system where the server becomes engaged in start-up work as the queue size, after an empty queue, builds up to $N(\geq 1)$. It may be seen that Eq. (6.4.35) also holds for $N = 0$, so that Eq. (6.4.32) can be obtained from (6.4.35) by putting $N = 0$.

(4) When SUT is zero, then $G_0^*(s) = 1$, so that

$$G_B^*(s) = [G^*(s)]^N$$

as can be expected.

(5) In particular, for the $M/M/1$ queue under N-policy and with exponential SUT with mean u, the LST of the busy period is given by

$$G_B^*(s) = \left[\frac{\beta(s)}{\rho}\right]^N \frac{1}{1 + u\{s + \lambda - \mu\beta(s)\}},$$

where

$$\beta(s) = \frac{(s + \lambda + \mu) \pm \sqrt{[(s + \lambda + \mu)^2 - 4\lambda\mu]}}{2\mu}.$$

6.5 Queues with Finite Input Source: $M/G/1//N$ System

Consider a situation like this. There are number of machines in an establishment that break down after being in operation for a random duration. A machine that breaks down is repaired by a single repairman, and when the repairman is busy repairing a machine, other broken machines form a queue in the service facility and wait for repair. Once repaired the machine starts working again and so on. A machine on an on-period (working) is said to be at *source*. Machines on an off-period (broken down) are said to be in the *service facility*, with one machine under repair with the repairman (server) and others waiting to get repaired (served). A machine is either at source (working) or at the service facility (failed). This is known as a machine interference problem.

Let N be the total number of machines. Assume that the lifetime of each machine is independent exponential with parameter λ—that is, the probability that a unit at the source arrives at the service facility during an infinitestimal interval Δt is $\lambda \Delta t$. Assume that the service time has a general distribution with DF $B(.)$, LST $B^*(.)$, and mean b. The model is denoted by $M/G/1//N$.

One is interested in the distribution of the queue size at the service facility as well as performance measures of the system in steady state. Denote

R: response time of a unit (queueing time plus service time of a unit) in the service facility

γ: throughput of the system (mean number of units served per unit time)

p_0: probability that the server is idle (no unit in the service facility) at an arbitrary time

L: number of units in the service facility at an arbitrary time

ρ: long-run fraction of time that the sever is busy $= 1 - p_0$

I: length of server idle period, and T: length of server busy period

$a = \lambda b$. Then

$$\gamma = \frac{\rho}{b} = \frac{1 - p_0}{b}, \tag{6.5.1}$$

also

$$\gamma = \frac{N}{E(R) + 1/\lambda}.$$

Thus,

$$E(R) = \frac{Nb}{1 - p_0} - \frac{1}{\lambda}.$$

Further, the arrival rate to the service facility equals the throughput (departure rate from it), so that

$$\gamma = \lambda[N - E(L)].$$

It follows that

$$E(L) = \gamma E(R)$$

(which is a relationship of the type of Little's Law).

Let π_0 be the probability that a busy period terminates after completion of service of a unit. (The unit is the last unit to be served in a busy period.) Then the mean number of units served during a busy period is $1/\pi_0$, and so $E(T) = b/\pi_0$. Using

$$p_0 = \frac{E(I)}{E(I) + E(T)},$$

$$\text{one gets} \quad p_0 = \frac{\pi_0}{\pi_0 + N\lambda b}, \tag{6.5.2}$$

so also γ, $E(T)$, $E(L)$ are expressible in terms of π_0.

One has to find the distribution $\{\pi_0, \pi_1, \ldots, \pi_{N-1}\}$, where π_k = probability that there are k units left behind in the service facility immediately after completion of service of a unit. Denote

L_n = number of units in the service facility immediately after service completion of nth unit, $n = 1, 2, \ldots$

The sequence $\{L_n, n = 1, 2, \ldots\}$ constitutes an embedded Markov chain having transition probability

$$p_{ij} = Pr\{L_n = j \mid L_{n-1} = i\}. \quad \text{Then}$$

$$p_{ij} = \begin{cases} \binom{N-1}{j} \int_0^\infty e^{-(N-1-j)\lambda x}(1 - e^{-\lambda x})^j \, dB(x), & i = 0 \\ \binom{N-i}{j-i+1} \int_0^\infty e^{-(N-1-j)\lambda x}(1 - e^{-\lambda x})^{j-i+1} \, dB(x), & j \geq i - 1 \geq 0 \\ 0, & i \geq 1, 0 \leq j \leq i - 1. \end{cases} \tag{6.5.3}$$

Now

$$\pi_j = \lim_{m\to\infty} p_{ij}^m = \lim_{m\to\infty} Pr\{L_{n+m} = j \mid L_n = i\}$$

exist for the irreducible, aperiodic Markov chain $\{L_n, n = 0, 1, \ldots, N-1\}$.

From the ergodic theorem of Markov chains (Theorem 1.1, Section 1.2.2.3), we see that π_j are given as solutions of

$$\begin{aligned} \pi_j &= \sum_{i=0}^{N-1} \pi_i p_{ij}, \quad 0 \le j \le N-1 \\ \text{and} \quad & \sum_{j=0}^{N-1} \pi_j = 1 \end{aligned} \tag{6.5.4}$$

(see also treatment of $M/G/1$ (Section 6.3.1)). Here, however, the pgf of $\{\pi_j\}$ cannot be put in an explicit form. An ingenious method is put forward to find $\{\pi_j\}$.

For details refer to Takagi (1993, Vol. II, Section 4.1).

It is found that π_0 is given by

$$\pi_0 = \frac{1}{\sum_{k=0}^{N-1} \binom{N-1}{k} \frac{1}{\zeta_k}},$$

where

$$\zeta_0 \equiv 1, \ \zeta_k \equiv \prod_{j=1}^{k} \frac{B^*(j\lambda)}{1 - B^*(j\lambda)}, \quad k = 1, 2, \ldots, N-1 \tag{6.5.5}$$

Once π_0 is found, one can find the performance measures as described above.

Note: Whereas, in our usual notation, λ is taken as the rate of arrival from an *infinite* source, here λ is taken as the rate of arrival of *each* unit from the source (to the service facility). Thus, here $E(I) = \frac{1}{N\lambda}$. The limiting case $N \to \infty$ is considered next.

Limiting Case: M/G/1 System

Taking limits as $N \to \infty, \lambda \to 0$ so that $N\lambda = \lambda'$ has a fixed finite value, one can find the expressions for the $M/G/1$ system. Taking limits, it can be seen that

$$\pi_0 = 1 - \rho + \frac{\lambda'}{N}\left[\frac{\lambda' b^{(2)}}{2(1-\rho)} + b\right] + O\left(\frac{1}{N}\right). \tag{6.5.6}$$

And the Pollaczek-Khinchin mean value formula for response time is given by

$$E(R) = \frac{\lambda' b^{(2)}}{2(1-\rho)} + b \tag{6.5.7}$$

(Takagi, 1993).

6.6 System with Limited Waiting Space: $M/G/1/K$ System

Here the service facility can accommodate at most K units, including the one in service, if any. For $K \to \infty$, we get the $M/G/1$ system with infinite space in the service facility.

Let P_B be the probability that an arriving unit is blocked because it finds the system full on arrival. The effective rate of arrival is thus $\lambda(1 - P_B)$, λ being the rate at which units arrive in the system. The utilization factor is

$$\begin{aligned} \rho &= \lambda(1 - P_B)b, \text{ where } b \text{ is the mean service time} \\ &= a(1 - P_B), \text{ where } a = \lambda b. \end{aligned} \tag{6.6.1}$$

This is the long-run fraction of time the server is busy at an arbitrary time. In this system, a unit may depart either on completion of service or on arrival if it finds the system full. Let us consider the system size at departure epochs (after service completion) and let L_n be the number of units left behind in the system immediately after completion of the nth service.

Then $\{L_n, n = 0, 1, 2, \ldots, K - 1\}$ is a Markov chain having transition probability

$$p_{jk} = Pr\{L_n = k \mid L_{n-1} = j\}. \tag{6.6.2}$$

Let

$$\pi_k = \lim_{m \to \infty} Pr\{L_{n+m} = k \mid L_n = j\} \tag{6.6.3}$$

and

$$\begin{aligned} a_k &= Pr\{k \text{ units arrive during a service time}\} \\ &= \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} dB(t), \quad k = 0, 1, 2, \ldots. \end{aligned}$$

Then

$$\begin{aligned} p_{0k} &= a_k, \quad 0 \le k \le K - 2 \\ &= \sum_{i=K-1}^{\infty} a_i, \quad k = K - 1 \end{aligned} \tag{6.6.4}$$

and for $1 \le j \le K - 1$

$$\begin{aligned} p_{jk} &= a_{k-j+1}, \quad j - 1 \le k \le K - 2 \\ &= \sum_{i=K-j}^{\infty} a_i, \quad k = K - 1. \end{aligned} \tag{6.6.5}$$

From the ergodic theorem of Markov chain, $\{\pi_k\}$ can be obtained as solutions of

$$\pi_k = \sum_{j=0}^{K-1} \pi_j p_{jk}, \quad 0 \le k \le K-1 \tag{6.6.6}$$

$$\text{and} \quad \sum_{k=0}^{K-1} \pi_k = 1. \tag{6.6.7}$$

There are $K+1$ equations for the K unknown probabilities $\pi_k, k = 0, \ldots, K-1$ so that one of the equations will be redundant (say, the one with $k = K-1$), so that we consider (6.6.6) for $k = 0, 1, 2, \ldots, K-2$.

So far, we have the same type of relations (though finite in number) as in the case of the $M/G/1$ system. However, a different method is needed to solve the equations (6.6.6), with (6.6.7) for $\pi_k, k = 0, 1, \ldots, K-1$.

An algorithmic method (Cooper, 1981) is as follows. Let

$$\pi_k' = \pi_k/\pi_0, \quad 0 \le k \le K-1 \tag{6.6.8}$$

Then

$$\pi_{k+1}' = \frac{1}{a_0}\left(\pi_k' - \sum_{j=1}^{k} \pi_j' a_{k-j+1} - a_k\right), \quad 0 \le k \le K-2. \tag{6.6.9}$$

These equations form a recursive system with $\pi_0' = 1$, so that $\pi_k', k = 1, 2, \ldots, K-1$ can be found. From (6.6.7) we get

$$\sum_{k=0}^{K-1} \pi_k' = \frac{1}{\pi_0}$$
$$\text{or} \quad \pi_0 = \left[\sum \pi_k'\right]^{-1}. \tag{6.6.10}$$

Thus, one can get $\pi_k, k = 0, 1, \ldots, K-1$.

Determination of arrival epoch distribution

Now let us find the probability distribution $\{P_k, \ 0 \le k \le K\}$, where $P_k = Pr\{\text{That there are } k \text{ units present in the system at an arbitrary point of time}\}$.

Let $\bar{\pi}_k = Pr\{\text{an arrival (whether he can join the queue or not) finds } k \text{ in the system}\}$. Since PASTA holds

$$\bar{\pi}_k = P_k, \quad 0 \le k \le K. \tag{6.6.11}$$

Since arrivals and departures occur one by one, we have (by Theorem 2.1), the probability distribution $\{\pi_k, \ 0 \le k \le K-1\}$ for the number in the system immediately after a service-completion is the same as the distribution

$\{\bar{\pi}_k,\ 0 \le k \le K-1\}$ of the number in the system immediately before an arrival, excluding those that find the system blocked. In other others $\bar{\pi}_k$ $(0 \le k \le K-1)$ is the conditional probability that an arriving unit find the system open (not blocked). The probability that the system is not blocked is $(1 - P_B)$. Thus,

$$\bar{\pi}_k = (1 - P_B)\pi_k, \quad 0 \le k \le K-1. \tag{6.6.12}$$

Thus, the arrival epoch distribution $\{\bar{\pi}_k,\ 0 \le k \le K-1\}$ can be obtained.

Distribution of system size at arbitrary time

P_k and $\pi_k, 0 \le k \le K-1$ satisfy the same set of equations (6.6.6) and (6.6.7). Thus, $P_k = C\pi_k (0 \le k \le K-1)$, where C is a constant to be determined. Also $\sum_{k=0}^{K} P_k = 1$ implies

$$C\sum_{k=0}^{K-1} \pi_k + P_K = 1,$$

$$\text{or} \quad P_K = 1 - C \text{ (using (6.6.7))}. \tag{6.6.13}$$

Now $P_K = \bar{\pi}_K = P_B$ is the blocking probability. From (6.6.1)

$$P_K = P_B = 1 - \rho/a = 1 - \frac{1 - P_0}{a} = 1 - \frac{1 - C\pi_0}{a}. \tag{6.6.14}$$

From (6.6.13) and (6.6.14) one gets

$$C = \frac{1}{\pi_0 + a} \tag{6.6.15}$$

$$P_k = C\pi_k = \frac{\pi_k}{\pi_0 + a} \quad 0 \le k \le K-1 \tag{6.6.16}$$

$$\text{and} \quad P_B = P_K = 1 - \frac{1}{\pi_0 + a}. \tag{6.6.17}$$

These are Cooper-Gebhardt relations (Gebhardt, 1973; Cooper, 1981; Fujiki and Gambe, 1980). Thus, $\{P_k,\ 0 \le k \le K\}$ is known.

One can now find the expected number in the system $\sum_{k=0}^{K} kP_k$ and also the expected response time in terms of π_k which could be determined algorithmically, as described above.

A more general system $M/G/1/K/N$, $K \le N$, with finite input source N and finite space K before the single server, has been considered by Takine *et al.* (1993).

For detailed treatment of these two finite systems, refer to Takagi (1993) and the references therein.

6.7 The $M^X/G/1$ Model with Bulk Arrival

6.7.1 The number in the system at departure epochs in steady state (Pollaczek-Khinchin formula)

Assume that the arrival epochs occur in accordance with a Poisson process with rate λ and the number of arrivals at each epoch is given by a **RV** X having distribution $a_j = Pr(X = j)$, and PGF

$$A(s) = \sum_j a_j s^j \quad \text{and} \quad E(X) = \sum j a_j = A'(1) = a. \tag{6.7.1}$$

$\{a_j\}$ is the batch size distribution.

The total arrivals A constitute a compound Poisson process having PGF $\exp\{-\lambda[1 - A(s)]\}$.

Suppose that N is the total number of arrivals during the service time of a customer. Then the PGF of N is given by

$$E[s^N] = K(s) = B^*[\lambda - \lambda A(s)], \tag{6.7.2}$$

where $B^*(s) = \int_0^\infty e^{-st}\, dB(t)$. (See Eq. (6.2.5a).)

The traffic intensity is $\rho = \lambda E(X)/\mu = \lambda a/\mu$. Assume that $\rho < 1$ so that the steady state is reached. The Pollaczek-Khinchin formula can now be extended for $M^X/G/1$. Writing the preceding expression of $K(s)$ in the Pollaczek-Khinchin formula given by Eq. (6.3.7) for $M/G/1$, we get the expression for the PGF $V(s)$ of the number in the system at departure epochs in steady state. $V(s)$ is then given by

$$V(s) = \frac{(1-\rho)(1-s)B^*(\lambda - \lambda A(s))}{B^*(\lambda - \lambda A(s)) - s}, \tag{6.7.3}$$

which is the Pollaczek-Khinchin formula for $M^X/G/1$.

In the particular case when $a_1 = 1, a_j = 0, j > 1$, we get $A(s) = s$ and $K(s) = B^*(\lambda - \lambda s)$, and we have an $M/G/1$ queue.

Note that $p_0 = 1 - \rho$ holds also for $M^X/G/1$ system.

6.7.2 Waiting-time distribution

Burke (1975) obtained the waiting-time distribution in an $M^X/G/1$ queueing system, thereby refining the results obtained by some authors earlier. Next we discuss Burke's approach.

Consider a test unit and let D be the total waiting time of the unit in queue—that is, D is the queueing time of an arbitrary test unit. The delay D is seen by the test unit to consist of two independent delays, D_1 and D_2. D_1 is the delay (or waiting time) of the first member to be served of the batch in which the test

unit arrives, and D_2 is the delay caused by the service times of the members of this batch that are served prior to the test unit—in other words, $D = D_1 + D_2$. Let W, W_i be the DF of D, D_i, $i = 1, 2$, respectively, and let $W^*(s)$, $W_i^*(s)$ be the LST of W, W_i, respectively. Let $B(t)$ be the service-time distribution and $B^*(s)$ be its LST. Denote $\beta^*(s)$ = LST of the DF of the total service time of all customers belonging to the same arrival group. Then

$$\begin{aligned}\beta^*(s) &= \sum_{k=1}^{\infty} a_k [B^*(s)]^k \\ &= A[B^*(s)]. \end{aligned} \tag{6.7.4}$$

To find the delay D_1, consider a batch as a whole as a single *supercustomer.* Then the LST of the waiting time of the first member of the batch in which the test unit arrives can be obtained from the corresponding expression of an $M/G/1$ system with $B^*(s)$ replaced by $\beta^*(s)$. That is, if $\rho = \lambda a/\mu < 1$, then replacing $B^*(s)$ by $\beta^*(s)$ (given in (6.5.4)) in the Pollaczek-Khinchin formula (6.3.11a), we get

$$\begin{aligned} W_1^*(s) &= \text{LST of the delay } D_1 \\ &= \frac{s(1-\rho)}{s - \lambda[1 - A(B^*(s))]}. \end{aligned} \tag{6.7.5}$$

Let p_i be the probability that the test customer arrives in a batch of size i. Let K be a significantly large number. Then in the first K batches of arrivals, the number of batches with i arrivals will be approximately $a_i K, i = 1, 2, \ldots$ and the total number of customers arriving in batches of size i will be approximately $i a_i K$. Thus, the total number of arrivals in K batches is

$$\sum_{i=1}^{\infty} i a_i K$$

and the proportion of those arriving in batches of size i is

$$\frac{i a_i K}{\sum_i i a_i K} = \frac{i a_i}{\sum_i i a_i} = \frac{i a_i}{a},$$

where $a = E(X)$.

Thus, for large K,

$$p_i = \frac{i a_i}{a}. \tag{6.7.6}$$

Assume now that the test customer arrives in a batch of size i. Assume further that service within members of any batch is in random order. Then the probability that the test customer chosen is the jth in the batch of i is $1/i$, $j = 1, 2, \ldots, i$. Again, if he/she is the jth customer to be taken for service, his/her delay (or waiting time in the queue) will be equal to the service time of $(j - 1)$ customers of

the batch (of size i) in which he/she arrives and who are served prior to him/her. Now conditioning on the size of the batch i on which the test customer arrives, we get

$$P(D_2 \leq t) = W_2(t) = \sum_{i=1}^{\infty} Pr\{\text{delay} \leq t \mid \text{he arrives in a batch of size } i\}\, p_i$$

$$= \sum_{i=1}^{\infty} \left[\sum_{j=1}^{i} B^{(j-1)*}(t) \frac{1}{i} \right] p_i, \tag{6.7.7}$$

where B^{k*} is the k-fold convolution of B with itself. Thus,

$$\begin{aligned} W_2^*(s) &= \text{LST of } W_2(t) \\ &= \sum_{i=1}^{\infty} \frac{p_i}{i} \left\{ \sum_{j=1}^{i} \text{LST of } B^{(j-1)*}(t) \right\} \\ &= \sum_{i=1}^{\infty} \frac{ia_i}{ia} \left\{ \sum_{j=1}^{i} [B^*(s)]^{j-1} \right\} \\ &= \sum_{i=1}^{\infty} \frac{a_i}{a} \frac{1 - [B^*(s)]^i}{1 - B^*(s)} \\ &= \frac{1 - A[B^*(s)]}{a[1 - B^*(s)]}. \end{aligned} \tag{6.7.8}$$

Since $D = D_1 + D_2$, the LST of the total delay D or waiting time in the queue of the test customer has the LST given by

$$\begin{aligned} W^*(s) &= W_1^*(s) W_2^*(s) \\ &= \frac{s(1-\rho)}{s - \lambda + \lambda A[B^*(s)]} \frac{1 - A[B^*(s)]}{a[1 - B^*(s)]}. \end{aligned} \tag{6.7.9}$$

The waiting time in the system or response time of the test unit is given by $W_s = D + v$, where v is the service time. Thus, the LST $W_s^*(s)$ of W_s, the response time, is given by

$$W_s^*(s) = W^*(s) B^*(s).$$

Notes:

(1) Burke obtains result (6.7.9) by taking into account the difference between the distribution of the size of a batch when sampled over batches and when sampled over units.

(2) While the simple proof just given of the crucial result $p_i = ia_i/a$ is due to Ross (1980), a rigorous proof using Blackwell's renewal theorem has been given by Burke.

(3) The number of other customers besides the test customer arriving in the same batch has the same Poisson distribution *iff* $\{a_i\}$ is Poisson. Similarly, if the test customer is equally likely to be in each of the positions in a batch, the number of customers *in front of him/her* has the same geometric distribution as a_j *iff* $\{a_i\}$ is geometric. (See Whitt, 1983).

6.7.2.1 Particular cases

(1) Single-arrival case: The Model $M/G/1$
Here

$$a_1 = 1, \quad a_i = 0, \quad i \neq 1, \quad a = 1, \quad A(s) = s \quad \text{and}$$
$$B^*(s) = A[B^*(s)] = B^*(s),$$

so that

$$W_Q^*(s) = W^*(s) = \frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)} \frac{1 - B^*(s)}{1 - B^*(s)}$$
$$= \frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)}.$$

If $G \equiv M$—that is, for $M/M/1$, $B^*(s) = \mu/(s+\mu)$ and the waiting time in the queue has LST

$$W_Q^*(s) = \frac{(1-\rho)(s+\mu)}{s + \mu - \lambda};$$

and the waiting time in the system has LST

$$W_s^*(s) = W_Q^*(s)\frac{\mu}{s+\mu} = \frac{(1-\rho)\mu}{s + (\mu - \lambda)}$$
$$= \frac{\mu - \lambda}{s + (\mu - \lambda)}$$

and is exponential with mean $1/(\mu - \lambda)$.
(2) The system $M^X/G/1$ where X is geometric having distribution

$$P(X = k) = pq^{k-1}, \quad k = 1, 2, \ldots, \quad 0 < p, q < 1,$$

then

$$a = 1/p,$$
$$A[B^*(s)] = \sum_{k=1}^{\infty} pq^{k-1}[B^*(s)]^k$$
$$= \frac{pB^*(s)}{1 - qB^*(s)},$$

and when $G \equiv M$—that is, $M^X/M/1$ with geometric arrival distribution $B^*(s) = \mu/(s+\mu)$, we get

$$A[B^*(s)] = \frac{p\mu}{s + p\mu}.$$

Thus, the waiting time in the queue has LST

$$W^*(s) = \frac{p(1-\rho)(s+\mu)}{s + (p\mu - \lambda)},$$

and the waiting time in the system has LST

$$\begin{aligned} W_s^*(s) &= W^*(s)B^*(s) \\ &= \frac{p\mu - \lambda}{s + (p\mu - \lambda)}; \end{aligned}$$

the waiting time in the system is exponential with mean $1/(p\mu - \lambda)$. This is another example of memoryless distributions leading to another memoryless distribution.

(3) $M^X/G/1$ *with a random setup time* V

Suppose that on arrival of the first unit after the termination of a busy period, the server has to set up the system for service, which requires a random setup time (denoted by RV V) before starting actual service. Then $I + V \equiv W$ is the time required for service to start after a busy period (I is the period during which the system is empty). Then $\{W\}$ and $\{B\}$ (busy period) define an alternating renewal process. We have

$$\frac{\lambda E(X)}{\mu} \equiv \rho = \frac{E(B)}{E(B) + E(W)} = \frac{E(B)}{E(B) + [1/\lambda + E(V)]},$$

whence we get

$$\begin{aligned} E(B) &= \frac{E(X)[1 + \lambda E(V)]}{\mu(1-\rho)} \\ &= \frac{1}{\mu(1-\rho)} \cdot K, \end{aligned}$$

where $K =$ average number of arrivals during W, the server idle period (during which there is no service).

It can be seen that

$$E(B) = \frac{\rho}{1-\rho} E(W)$$

holds, where W and B are server idle and server busy periods, respectively.

6.7.2.2 Moments of $D = D_1 + D_2$ for $M^X/G/1$

We have

$$E(D_1) = E(D_1) + E(D_2)$$

$$E(D_1) = -\frac{d}{ds} W_1^*(s)|_{s=0} = -\frac{d}{ds} \frac{s(1-\rho)}{s - \lambda + \lambda A[B^*(s)]}\bigg|_{s=0}$$

$$= -(1-\rho) \frac{s - \lambda + \lambda A(B^*(s)) - s\left\{1 + \lambda A'(B^*(s)) \frac{d}{ds} B^*(s)\right\}}{\{s - \lambda + \lambda A[B^*(s)]\}^2}\bigg|_{s=0}$$

(which is of the form 0/0). Using L'Hôpital's rule and simplifying, we get

$$E(D_1) = \frac{\lambda(1-\rho)}{2(1-\rho)^2} \frac{d^2}{ds^2} A[B^*(s)]\,|_{s=0},$$

where $(d^2/ds^2)\,A[B^*(s)]$ is the second moment of the supercustomer's service time. On simplification, we get

$$E(D_1) = \frac{\lambda}{2(1-\rho)}\{A''[B^*(s)]\}\left[\frac{d}{ds} B^*(s)\right]^2 + A'[B^*(s)] \frac{d^2}{ds^2} B^*(s)|_{s=0}$$

$$= \frac{\lambda}{2(1-\rho)}\left[A''(1)\left(-\frac{1}{\mu}\right)^2 + A'(1)\mu_2\right], \tag{6.7.10}$$

where $\mu_i = E(v^i)$. Writing $a^{(2)} = E(X^2)$, we get $A''(1) = E(X^2) - E(X) = a^{(2)} - a$. Thus,

$$E(D_1) = \frac{\lambda}{2(1-\rho)}\left[\frac{a^{(2)} - a}{\mu^2} + a\mu_2\right]. \tag{6.7.11}$$

Again,

$$E(D_2) = -\frac{d}{ds} W_2^*(s)|_{s=0} \quad \text{or}$$

$$aE(D_2) = -\frac{d}{ds} \frac{1 - A[B^*(s)]}{[1 - B^*(s)]}\bigg|_{s=0}$$

$$= \frac{\left\{A'[B^*(s)] \frac{d}{ds} B^*(s)\right\}[1 - B^*(s)] - \{1 - A[B^*(s)]\} \frac{d}{ds} B^*(s)}{[1 - B^*(s)]^2}\bigg|_{s=0}$$

(which is of the form 0/0). Using L'Hôpital's rule and simplifying, one gets

$$E(D_2) = \frac{1}{2\mu}\left(\frac{a^{(2)}}{a} - 1\right). \tag{6.7.12}$$

Thus,

$$E(D) = \frac{1}{2(1-\rho)}\left(\frac{a^{(2)}-a}{\mu^2} + a\mu_2\right) + \left(\frac{a^{(2)}}{a} - 1\right)\frac{1}{2\mu}. \quad (6.7.13)$$

Alternative form of E(D)

Writing $\mu_2 = (1 + c_s^2)/\mu^2$, where c_s = coeff. of variation of the service time S. We get from (6.7.13), on simplification,

$$E(D) = \frac{\rho}{2\mu(1-\rho)}\{1 + c_s^2\} + \frac{a^{(2)}-a}{a}\left\{\frac{1}{2\mu(1-\rho)}\right\}. \quad (6.7.13a)$$

The first term is the $E(W_q)$ for standard $M/G/1$ queue, here with $\rho = \lambda E(X)/\mu$ (see Eq. (6.3.12)), and the second term reflects the additional effect of the batch size.

The average number of customers L_Q in the queue is thus given by

$$L_Q = \lambda E(X) E(D). \quad (6.7.14)$$

This can also be found from Eq. (6.7.3).

Again the first term of Eq. (6.7.13a)

$$= \frac{\rho}{1-\rho} \cdot \frac{E(S^2)}{2E(S)} = \frac{\rho}{1-\rho} E(S_R),$$

where S_R = residual service time of the unit in service at the instant of arrival of the test customer.

The second term of Eq. (6.7.13a)

$$= \frac{E(S)}{1-\rho}\left[\frac{1}{2}\left(\frac{E(X^2)}{E(X)} - 1\right)\right]$$
$$= \frac{E(S)}{1-\rho} E(X_R),$$

where X_R = residual group size (of the batch in which the test unit arrives) and who are served prior to the test unit. Thus,

$$E(D) = \frac{\rho}{1-\rho} E(S_R) + \frac{E(S)}{1-\rho} E(X_R). \quad (6.7.15)$$

This is an insightful result.

In the case of $M/G/1$ queue, $E(X_R) = 0$. (See Chae and Lee (1995).)

6.7.3 Feedback queues

Feedback queues relate to those queues in which a customer served once, when his service becomes unsuccessful, and is served again and again till his service becomes successful (see, for example, Takács (1963), Kleinrock (1975), Takagi (1991), and Boxma and Yechiali (1997)). Many real-life situations could be modeled as a feedback queue. For example, in data transmission, a packet transmitted from the source to the destination may be returned and it may go on like that until the packet is finally transmitted.

Let us assume that units (customers/messages/packets) arrive singly in a Poisson stream with rate λ and that the service time B is general with DF $B(.)$ and LST $B^*(.)$. As soon as the service is completed, a unit, whose service is successful departs from the system with probability v, or if the service is unsuccessful, the unit is cycled back into the system with probability $1 - v$. This is known as Bernoulli Feedback, and we shall denote the system by $M/G/1$ (BF). Now there can be different ways as to how the unsuccessful unit is admitted into the system. In Takács's model, such a unit joins the tail of the queue; in Takagi's model (1996), the unit joins the queue where the service discipline is SIRO (service in random order).

Here we consider the simple case where the unsuccessful unit joins at the head the queue and is immediately taken for service again and again.

For a feedback queue, the response time R—that is, waiting time W in the system from the arrival of the unit till it finally departs from the system—is more meaningful. The expected response time is the same for all the different service disciplines.

The total service time B_T of a unit is the time from the instant of commencement of his first service to the instant of its final departure from the system, with no or at least one feedback with the same probability $1 - v$ of feedback after each service. The LST of B_T given by

$$\begin{aligned} B_T^*(s) &= vB^*(s) + (1-v)v[B^*(s)]^2 + (1-v)^2v[B^*(s)]^3 + \cdots \\ &= \frac{vB^*(s)}{1-(1-v)B^*(s)}. \end{aligned} \tag{6.7.16}$$

Its moments, when these and those of B exist, satisfy the recurrence relation

$$b_T^{(r)} = b^{(r)} + \frac{1-v}{v}\sum_{i=1}^{r}\binom{r}{i}b^{(i)}b_T^{(r-i)} \quad r = 1,2,3,\ldots, \tag{6.7.17}$$

where

$$b_T^{(r)} = E\left[B_T^r\right], b^{(r)} = E(B^r), b^{(1)} = b, b^{(0)} = 1$$

$$b_T^{(1)} = b_T, b_T^{(0)} = 1.$$

One gets

$$b_T^{(1)} \equiv b_T = \frac{b}{v}$$

and

$$b_T^{(2)} = \frac{b^{(2)}}{v} + \frac{2(1-v)}{v^2} b^2.$$

Thus,

$$\rho = \lambda b / v.$$

The PGF $V(z)$ of the system size at departure points can be obtained from the Pollaczek-Khinchin formula replacing $B^*(s)$ by $B_T^*(s)$.

The system size (and queue size) does not depend on the service discipline. The system size in an $M/G/1$ (BF) system is equivalent to the number of batches present in a Poisson batch arrival system $M^X/G/1$ without feedback, the random batch size X being a zero-truncated geometric having PMF

$$Pr\{X = k\} = v(1-v)^{k-1}, \quad k = 1, 2, \ldots$$

and PGF

$$A(z) = \sum_{k=1}^{\infty} Pr\{X = k\} z^k = \frac{vz}{1-(1-v)z}. \tag{6.7.18}$$

It follows that the delay (queueing time) D of a test unit in an $M/G/1$ (BF) queue is equal to the delay of the first member (to be taken for service) of a (test) batch in an $M^X/G/1$ queue without feedback. Thus the LST of D is given by

$$D^*(s) = \frac{s(1-\rho)}{s - \lambda\{1 - B_T^*(s)\}}. \tag{6.7.19}$$

Hence, the LST of the response time R in an $M/G/1$ (BF) queue is given by

$$R^*(s) = D^*(s) B_T^*(s).$$

We have

$$E\{R\} = E\{D\} + E\{B_T\} = \frac{2(1-\lambda b)b + \lambda b^{(2)}}{2(v - \lambda b)}. \tag{6.7.20}$$

The expected $E(R)$ is the same for all the disciplines, and this is a more convenient way to find it. Medhi (2001) finds the $E(R^2)$ for the discipline considered here and gives some numerical comparisons. The extension of the model to cover server's vacation and threshold policy is also considered. Extension to a more general model as examined by Boxma and Yechiali (1997), where the

service time of the unit taken for its first service is different from the subsequent service times of unsuccessful units, is also indicated.

6.8 The $M/G\ (a, b)/1$ Model with General Bulk Service

Let A = number of arrivals during the service time of a unit and

$$q_r = Pr\{A = r\}, \quad r = 0, 1, 2, \ldots.$$

Then A has the PGF

$$Q(s) = \sum_r q_r s^r = B^*(\lambda - \lambda s),$$

where $B^*(s)$ is the LST of $B(t)$. It can be easily seen that $\{X_n, n \geq 0\}$ (where $X_n = N(t_n + 0)$, the number left behind by the batch departing at instant t_n) is a Markov chain.

The transition probabilities

$$p_{ij} = Pr\{X_{n+1} = j \mid X_n = i\}$$

of the chain $\{X_n, n \geq 0\}$ can be written in terms of q_r as follows:

$$\begin{aligned} p_{ij} &= q_j, && 0 \leq i \leq b, \quad j \geq 0 \\ &= q_{j-i+b}, && i \geq b, \quad j \geq i - b \\ &= 0, \text{ in all other cases.} \end{aligned} \tag{6.8.1}$$

The TPM of the denumerable Markov chain can be written as follows.

$$\mathbf{P} = (p_{ij}) \equiv \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ b \\ b+1 \\ b+2 \\ \cdots \\ \cdots \\ \cdots \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & \cdots \end{array} \\ \begin{bmatrix} q_0 & q_1 & q_2 & q_3 & \cdots \\ q_0 & q_1 & q_2 & q_3 & \cdots \\ q_0 & q_1 & q_2 & q_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_0 & q_1 & q_2 & q_3 & \cdots \\ 0 & q_0 & q_1 & q_2 & \cdots \\ 0 & 0 & q_0 & q_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \end{bmatrix} \end{array} \tag{6.8.2}$$

The chain is irreducible and aperiodic. It can be shown that it is persistent non-null when $\rho = (\lambda/b\mu) < 1$. It follows that probabilities

$$v_j = \lim_{n\to\infty} p_{ij}^{(n)}, \quad j = 0, 1, 2, \ldots$$

exist and $\mathbf{V} = (v_0, v_1, \ldots)$ is given as the unique solution of $\mathbf{V} = \mathbf{VP}$. We have

$$v_j = \left(\sum_{r=0}^{b-1} v_r\right) q_j + \left(\sum_{r=0}^{j} q_{j-r} v_{b+r}\right), \quad j \geq 0.$$

The PGF of $\{v_j\}$ is given by

$$\begin{aligned} V(s) &= \sum_{j=0}^{\infty} v_j s^j \\ &= \left(\sum_{r=0}^{b-1} v_r\right) Q(s) + \sum_{r=0}^{\infty} \left\{\sum_{j=r}^{\infty} s^{j-r}\right\} v_{b+r} s^r \\ &= \left(\sum_{r=0}^{b-1} v_r\right) Q(s) + Q(s) \sum_{r=0}^{\infty} v_{b+r} s^r \\ &= \left(\sum_{r=0}^{b-1} v_r\right) Q(s) + \left(\frac{Q(s)}{s^b}\right)\left[V(s) - \sum_{r=0}^{b-1} v_r s^r\right]. \end{aligned}$$

On simplification, we get

$$V(s) = \frac{\sum_{r=0}^{b-1}(s^r - s^b) v_r}{Q(s) - s^b}. \tag{6.8.3}$$

Putting $B^*(\lambda - \lambda s)$ for $Q(s)$, we get

$$V(s) = \frac{\sum_{r=0}^{b-1}(s^r - s^b) v_r}{B^*(\lambda - \lambda s) - s^b} B^*(\lambda - \lambda s). \tag{6.8.4}$$

Notes:

(1) For the standard $M/G/1$ model, $a = b = 1$. Then $v_0 = p_0$ (as PASTA holds) and $p_0 = 1 - \rho$ (for a single-server model), so that $v_0 = 1 - \rho$. We can get the Pollaczek-Khinchin formula (as given in (6.3.7a)) from the above formula (6.8.4). Thus, (6.8.4) can be considered as an extension of the Pollaczek-Khinchin formula.

(2) The distribution $\{v_j\}$ is independent of a and holds for $a = 1, 2, \ldots$ and even for $a = 0$.

(3) The transient distribution of $N(t)$ has been studied, among others, by Neuts (1967), who also obtains π_j, where

$$\pi_j = \lim_{t\to\infty}\{N(t) = j \mid N(0) = i\}.$$

(4) Borthakur (1975) discusses the busy period distribution.

(5) More general models have also been studied: Borthakur and Medhi (1974) examine transient distribution of $M^X/G(a,b)/1$. Jacob *et al.* (1988) consider transient solutions of $M/G(a,b)/1/N$ with finite buffer. This model was also studied by Gold and Tran-Gia (1993) and later by another method by Chaudhry and Gupta (1999).

Bertsimas and Papaconstantinou (1988) consider a model $M/C_2(a,b)/s$ (having bulk service and Coxian service-time distribution) of importance in transportation.

6.9 The G/M/1 Model

6.9.1 Steady-state arrival epoch system size

Let

$$\begin{aligned} u &= \text{interarrival time} \\ A(t) &= Pr\{u \geq t\} \\ A^*(s) &= \text{LST of } u \end{aligned}$$

and $t_n, n = 1, 2, \ldots (t_0 = 0)$ be the instant at which the nth arrival occurs (or nth unit arrives). Then $N(t_n - 0) = Y_n, n = 0, 1, 2, \ldots$ gives the number in the system immediately preceding the nth arrival. We have

$$Y_{n+1} = Y_n + 1 - B_{n+1}, \quad \text{if} \quad Y_n \geq 0, \quad B_{n+1} \leq Y_n + 1, \qquad (6.9.1)$$

where B_{n+1} is the number of units served during $(t_{n+1} - t_n)$—that is, the interarrival time between the nth and $(n+1)$th arrivals. Clearly, $\{Y_n, n \geq 0\}$ is a denumerable Markov chain.

Now $B_n = B$ and

$$g_r = Pr\{B = r\} = \int_0^\infty \frac{e^{-\mu t}(\mu t)^r}{r!} dA(t) \quad r = 0, 1, 2, \ldots. \qquad (6.9.2)$$

The PGF of A is given by

$$\begin{aligned} G(s) = \sum g_r s^r &= \sum_r \int_0^\infty \frac{e^{-\mu t}(\mu t)^r}{r!} s^r dA(t) = \int_0^\infty e^{-\mu t} e^{s\mu t}\, dA(t) \\ &= A^*(\mu - s\mu). \end{aligned} \qquad (6.9.3)$$

The transition probabilities of the chain, denoted by

$$p_{ij} = Pr\{Y_{n+1} = j \mid Y_n = i\} \tag{6.9.4}$$

can be expressed in terms of the g_r's, as follows:

$$\begin{aligned} p_{ij} &= g_{i+1-j}, \quad i+1 \geq j \geq 1, \quad i \geq 0 \\ &= 0, \quad i+1 < j. \end{aligned} \tag{6.9.5}$$

For

$$j = 0,\ p_{ij} = 1 - \sum_{r=0}^{i} g_r = h_j \quad \text{(say)}. \tag{6.9.6}$$

The TPM of the chain can be put as

$$\mathbf{P} = (p_{ij}) = \begin{bmatrix} h_0 & g_0 & 0 & 0 & 0 & \dots \\ h_1 & g_1 & g_0 & 0 & 0 & \dots \\ h_2 & g_2 & g_1 & g_0 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{6.9.7}$$

The chain was considered in Example 1.4, Chapter 1. The chain is irreducible and aperiodic, and is persistent non-null when $\rho < 1$. Thus, when $\rho < 1$, the limiting arrival epoch system size probabilities

$$v_j = \lim_{n \to \infty} p_{ij}^{(n)} \tag{6.9.8}$$

exist and are given as the unique solution of

$$\boldsymbol{V} = \boldsymbol{V}\boldsymbol{P}, \tag{6.9.9}$$

$$\text{where} \quad \boldsymbol{V} = (v_0, v_1, \dots), \quad \sum v_j = 1.$$

Denote by r_0 the unique root inside $|z| = 1$ of $r(z) = z - A^*(\mu - \mu z) = 0$. We have

$$v_j = (1 - r_0) r_0^j, \quad j \geq 0$$

as a unique solution of Eq. (6.9.9).

6.9.1.1 Alternative method of finding V

We consider an alternative method of finding the unique solution of Eq. (6.9.9). The equations

$$v_j = \sum_i v_i p_{ij}, \quad j \geq 0, \quad \text{and} \quad \sum v_j = 1$$

reduce to

$$v_j = \sum_{i=j-1}^{\infty} v_i g_{i+1-j}$$

$$= \sum_{i=j-1}^{\infty} v_i \int_0^{\infty} \frac{e^{-\mu t}(\mu t)^{i+1-j}}{(i+1-j)!}\, dA(t), \quad j \geq 1, \quad (6.9.10)$$

and $\sum v_j = 1.$

As one equation is redundant, the equation for $j = 0$ is excluded.

Let us try a solution of the form $v_j = c\alpha^j$. Substituting in Eq. (6.9.9), we get

$$c\alpha^j = \sum_{i=j-1}^{\infty} c\alpha^i \int_0^{\infty} \frac{e^{-\mu t}(\mu t)^{i+1-j}}{(i+1-j)!}\, dA(t)$$

$$= c \int_0^{\infty} e^{-\mu t} \left[\sum_{i=j-1}^{\infty} \frac{(\mu t)^{i+1-j}}{(i+1-j)!} \alpha^i \right] dA(t)$$

(interchanging the summation and integration operators). Now

$$\sum_{i=j-1}^{\infty} \frac{(\mu t)^{i+1-j}}{(i+1-j)!}\alpha^i = \alpha^{j-1} \sum_{i=j-1}^{\infty} \frac{(\alpha\mu t)^{i+1-j}}{(i+1-j)!}$$

$$= \alpha^{j-1} \sum_{i=j-1}^{\infty} \frac{(\alpha\mu t)^k}{k!}$$

$$= \alpha^{j-1} e^{\alpha\mu t}.$$

Thus, from (6.9.10) we get

$$\alpha^j = \alpha^{j-1} \int_0^{\infty} e^{-\mu t} e^{\alpha\mu t}\, dA(t)$$

$$= \alpha^{j-1} \int_0^{\infty} e^{-\mu t(1-\alpha)}\, dA(t) \quad \text{or}$$

$$\alpha = A^*(\mu - \mu\alpha).$$

That is, α is a root of the equation

$$r(z) = z - A^*(\mu - \mu z) = 0.$$

From $\sum v_j = 1$, it follows that the root is of modulus less than 1 and that the constant is $c = 1 - \alpha$. As v_js are the unique solutions of (6.9.9), we get

$$v_j = (1-\alpha)\alpha^j, \quad j \geq 0.$$

α being the root inside $|z| = 1$ of $r(z) = 0$. It can be shown that when $\rho < 1$, there is a unique root inside $|z| = 1$. Thus, we have $\alpha = r_0$, so that

$$v_j = (1 - r_0)r_0^j, \quad j \geq 0. \tag{6.9.11}$$

Note: We have $v_j = (1 - r_0)r_0^j \quad j = 0, 1, 2, \ldots$. The distribution is geometric, having mean $r_0/(1 - r_0)$. An explanation for the occurrence of geometric distribution as a steady-state system size distribution of the $G/M/1$ queue has been put forward by Kingman (1963). An analytical proof of Eq. (6.9.11) via the Wiener-Hopf technique has been given by Neuts (1966).

Example 6.4. The $E_k/M/1$ model. Here $A^*(s) = (\lambda k/(s + \lambda k))^k$ and the characteristic equation reduces to

$$\begin{aligned} r(z) &\equiv z - A^*(\mu - \mu z) \\ &= z - \left(\frac{\lambda k}{\mu - \mu z + \lambda k}\right)^k = 0. \end{aligned} \tag{6.9.12}$$

In particular, for the $M/M/1$ model, $k = 1$, so that Eq. (6.9.12) reduces to a quadratic $z^2 - (1 + \rho)z + \rho = 0$ with $r_0 = \rho$ as the unique root inside $|z| = 1$. We thus have,

$$v_n = (1 - \rho)\rho^n, \quad n \geq 0$$

and as PASTA holds

$$p_n = v_n = (1 - \rho)\rho^n, \quad n \geq 0.$$

The $D/M/1$ model can be covered by taking the limit of $A^*(s)$ as $k \to \infty$. We have then $A^*(s) = e^{-s/\lambda}$, so that the characteristic equation reduces to

$$ze^{(1-z)/\rho} = 1.$$

Remarks: The contrasting situation between $M/G/1$ and $G/M/1$ systems may be described as follows. Unlike the $G/M/1$ queue, the equilibrium queue length and delay distributions of the $M/G/1$ queue *completely* characterize the arrival and service processes, of course, to a scale factor. For an $M/G/1$ system, it is possible to reconstruct the entire model from the equilibrium distribution and thence also to calculate any desired fluctuations in the queue-length process.

6.9.2 General time system size in steady state

By considering the embedded Markov chain $\{Y_n, n \geq 0\}$ (where Y_n is the system size immediately preceding the nth arrival), we obtain the limiting probabilities

$$v_j = \lim_{n \to \infty} p_{ij}^{(n)}.$$

For a fixed j, v_j is the probability that an arrival finds j in the system. We denote this probability by a_j, so that $v_j = a_j$.

Let $N(t)$ denote the system size at an arbitrary (general) time t, and let

$$p_{ij}(t) = Pr\{N(t) = j \mid N(0) = i\}, \tag{6.9.13}$$

then

$$\lim_{t\to\infty} p_{ij}(t) = p_j$$

(when it exists) gives the probability that there are j in the system in steady state. Further, p_j equals the long-run proportion of times that there are j customers in the system—that is, p_j are time averages. Since the arrival process is not Poisson, we cannot conclude that $a_j = p_j$ holds (for an $M/G/1$ system, the equality holds). Thus, a_j and p_j are different for this $G/M/1$ system, and we have to obtain the relation (if any) that exists between them. Takács (1962) obtains a relation between them. We give below Ross's (1980) probabilistic derivation. To find p_j, we first observe that the rate at which the number in the system changes from $(j-1)$ to j must be equal to the rate at which it changes from j to $(j-1)$. The rate of change from state $(j-1)$ to state j is equal to λ multiplied by the proportion of arrivals who find $(j-1)$ in the system. In other words, rate of change from state $(j-1)$ to state j equals λa_{j-1}, $j \geq 1$. Now the rate at which the system changes from state j to $(j-1)$ is equal to the proportion of time that there are j in the system multiplied by the rate of service μ. In other words, rate of change from state j to state $j-1$ equals μp_j, $j \geq 1$. Since these two rates are equal, we have

$$\lambda a_{j-1} = \mu p_j, \quad j \geq 1,$$

so that

$$\begin{aligned} p_j &= \left(\frac{\lambda}{\mu}\right) a_{j-1}, \; j \geq 1 \\ &= \rho v_{j-1} \text{ (since } a_j = v_j). \end{aligned} \tag{6.9.14a}$$

Thus,

$$\begin{aligned} p_j &= \left(\frac{\lambda}{\mu}\right)(1 - r_0) r_0^{j-1}, \quad j \geq 1 \\ &= \rho(1 - r_0) r_0^{j-1}, \quad j \geq 1. \end{aligned}$$

Since $\sum_{j=0}^{\infty} p_j = 1$, we have $1 = p_0 + \rho(1 - r_0)/(1 - r_0)$ so that $p_0 = 1 - \rho$, and

$$p_j = \begin{cases} 1 - \rho & j = 0 \\ \rho(1 - r_0) r_0^{j-1}, & j = 1, 2, \ldots, \end{cases} \tag{6.9.14b}$$

where r_0 is the unique root inside $|z| = 1$ of the characteristic equation $z - A^*(\mu - \mu z) = 0$. We have

$$E(N) = \frac{\rho}{1 - r_0} \tag{6.9.15}$$

and

$$\text{var}(N) = \frac{\rho(1 + r_0 - \rho)}{(1 - r_0)^2}.$$

In particular, for an $M/M/1$ model, $r_0 = \rho$, so that $p_j = (1 - \rho)\rho^j$, $j \geq 0$. For a more general relation between p_j and a_j, see Fakinos (1982).

6.9.2.1 System size at most recent arrival

Besides the two stochastic processes $\{Y_n, n \geq 0\}$ and $\{N(t), t \geq 0\}$ another related stochastic process $\{Z(t), t \geq 0\}$, where $Z(t) = Y_n, t_n \leq t < t_{n+1}$, may be considered. $Z(t)$ denotes the system size at the most recent arrival. $\{Z(t), t \geq 0\}$ is a semi-Markov process having $\{Y_n, n \geq 0\}$ for its embedded Markov chain, the transitions occurring at the arrival epochs. Let

$$f_{ij}(t) = Pr\{Z(t) = j \mid Z(0) = i\}. \tag{6.9.16}$$

Then

$$\lim_{t \to \infty} f_{ij}(t) = f_j$$

(when it exists) gives the limiting probability that the system size at the most recent arrival is j. Now v_j is the limiting probability (associated with the embedded Markov chain Y_n) that an arrival finds j in the system—that is,

$$v_j = \lim_{n \to \infty} p_{ij}^{(n)}.$$

We have found a relationship between v_j and p_j (as given in Eq. (6.9.14a)). Let us find a relationship, if any, existing between f_j and v_j (or p_j).

From the theory of semi-Markov process, we have

$$f_j = \frac{v_j m_j}{\sum_i v_i m_i}, \tag{6.9.17}$$

where m_i is the expected time spent in the state i during each visit. Here $m_i = 1/\lambda$ for all i, as the expected sojourn time in any state i is the same as the expected interarrival time, transitions occurring only at the arrival points (which are the regeneration points). Thus,

$$f_j = \frac{v_j/\lambda}{\sum_i v_i/\lambda} = \frac{v_j}{\sum_i v_i} = v_j \text{ for all } j(\geq 0), \tag{6.9.18}$$

as is to be expected.

6.9.3 Waiting-time distribution

As the distribution of $\{v_n\}$ is geometric and the service time is exponential, the waiting time in the queue W_q being random geometric sum of exponential

variables is modified exponential. We have $Pr\{W_q = 0\} = 1 - r_0$ and

$$Pr\{W_q \leq t\} = 1 - r_0 e^{-\mu t(1-r_0)t}, \quad t > 0. \tag{6.9.19}$$

We at once get $E\{W_q\}$ from Eq. (6.9.19). Alternatively,

$$\begin{aligned} E\{W_q\} &= \sum_{k=0}^{\infty} E\{\text{time in queue} \mid \text{an arrival finds } k \text{ in system}\} \\ &\quad \times Pr\{\text{an arrival finds } k \text{ in system}\} \qquad (6.9.20a) \\ &= \sum_{k=0}^{\infty} \frac{k}{\mu}(1 - r_0) r_0^k = \frac{r_0}{\mu(1 - r_0)}, \end{aligned}$$

and using Little's formula, we get

$$\{L_Q\} = \lambda E\{W_q\} = \frac{\lambda r_0}{\mu(1 - r_0)}. \tag{6.9.20b}$$

The conditional distribution of waiting time in the queue W_q given that a unit has to wait is, again, exponential with mean $1/\mu(1-r_0)$. For an $M/G/1$ system, we have from Eq. (6.3.12)

$$E\{W_q\} = \frac{\lambda}{2(1 - \rho)} \left\{\sigma^2 + \frac{1}{\mu^2}\right\},$$

and it follows that for a Poisson input queue with specified arrival and service rates, $E\{W_q\}$ is the least when $G \equiv D$—that is, when the service time is deterministic.

A similar result holds for the $G/M/1$ queue. This is stated as a folk theorem, which is as follows.

Folk Theorem 6.2. *Among all arrival processes for a single-server queue having exponential service time and with given arrival and service rates, the arrival process for which the average waiting time is minimum (and all moments of waiting time as well as other related quantities) is the process with constant (or deterministic) interarrival times.*

For proof of this theorem, see Hajek (1983) and Humblet (1982). They also show that the average waiting time in a $G/G/1$ queue is minimized by deterministic service time.

Note: The proof given by Hajek relies heavily on the convexity property of systems with exponential service-time distribution. The property that is of independent interest in itself is as follows.

The expected queue length in a system with exponential service time at a given time is a convex function of the set of previous interarrival times.

The result holds for higher moments of queue length as well.

Another related result obtained by Fischer (1974) is as follows. The average waiting time in steady state of an arriving customer decreases monotonically in k for a fixed arrival and service rates for both $M/E_k/1$ and $E_k/M/1$ systems. As regards the average waiting time for an exponential server queue with specified arrival and service rates, the queues with Poisson input and with deterministic input have, respectively, the greatest and the least average waiting times, among all Erlangian-type arrival processes. The folk theorem implies that the average waiting time is the least for deterministic input among all arrival processes.

6.9.4 Expected duration of busy period and idle period

Ross (1980) gives a simple and interesting method of finding the average duration of busy and idle periods of a $G/M/1$ system.

We have so far assumed that the queue discipline is FCFS (first-come/first-served). Even if the service discipline is not FCFS but LCFS (last-come/first-served) or service in random order, the distribution of the number of customers would be the same and so also would be that of the lengths of the busy and idle periods. However, the waiting time in the queue will have a different distribution, even though the average waiting time will be the same irrespective of the type of the discipline.

We assume that the queue discipline is LCFS, and we denote by $W_q^{(L)}$ the waiting time under LCFS. An arriving test customer finds the server idle with probability $a_0 = v_0 = 1 - r_0$ or busy with probability r_0. If he finds the server busy, he will have to wait

(i) until the remaining service time of the customer under service (which has the same exponential distribution as the full service time), and also

(ii) until the completion of service of each of the customers who arrives after him (service discipline being LCFS).

Now the customer under service can be thought of as initiating a server busy period B (its duration being the interval between the epoch of arrival of the test customer and the epoch of completion of services of all the customers who arrive after the test customer). Thus, if the test customer finds the server busy, he will have to wait in the queue for the same duration as the busy period B of the server. Conditioning on the state (idle or busy state of the server) in which the test customer finds the system on arrival, we get

$$\begin{aligned} E\{W_q^{(L)}\} &= E\{\text{wait} \mid \text{arrival finds server idle}\} \\ &\quad \times Pr\{\text{arrival finds none in system}\} \\ &\quad + E\{\text{wait} \mid \text{arrival finds server busy}\} \\ &\quad \times Pr\{\text{arrival finds server busy}\} \\ &= 0 \times a_0 + E\{B\}[(1 - a_0)] \\ &= E(B)r_0, \end{aligned}$$

but

$$E\{W_q^{(L)}\} = E\{W_q\} = \frac{r_0}{\mu(1-r_0)} \quad \text{(from (6.7.20}a\text{))}.$$

Thus,

$$E(B) = \frac{1}{\mu(1-r_0)}, \tag{6.9.21}$$

and the average number served during a busy period is $\frac{1}{1-r_0} = \frac{1}{a_0}$.

Again, we have

$$\frac{E(I)}{E(I)+E(B)} = p_0 = 1-\rho, \tag{6.9.22}$$

where I is the idle period so that we get

$$E(I) = \frac{1-\rho}{\lambda(1-r_0)}.$$

Notes:

(1) For Poisson input queue $r_0 = \rho$; we have then the results for an $M/M/1$ queue.

(2) Laxmi and Gupta (1999) discuss $GI/M^{[b]}/1/N$, queue with finite buffer and batch service of maximum size b.

(3) Brandt (1987) deals with $G/M/s/r$ model.

6.10 Multiserver Model

6.10.1 The $M/G/\infty$ model: transient-state distribution

Let

$N(t)$ = number in the system (or number of busy channels) at epoch t
$A(t)$ = number of arrivals by time t—that is, in the interval $(0, t)$
$D(t)$ = number of departures by time t
$p_n(t) = Pr\{N(t) = n\}$

so that

$$A(t) = N(t) + D(t).$$

Assume that $p_o(0) = 1$, $p_n(0) = 0, n \neq 0$. We have, by conditioning on $A(t)$,

$$\begin{aligned} p_n(t) &= Pr\{N(t) = n\} \\ &= \sum_{k=n}^{\infty} Pr\{N(t) = n \mid A(t) = k\} Pr\{A(t) = k\} \\ &= \sum_{k=n}^{\infty} Pr\{N(t) = n \mid A(t) = k\} \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \end{aligned} \tag{6.10.1}$$

since the arrival process is Poisson. Now,

$$\begin{aligned} Pr\{N(t) = n \mid A(t) = k\} &= Pr\{\text{out of } k \text{ arrivals in } (0, t) \text{ the number} \\ &\quad \text{still in service at epoch } t \text{ is } n\} \\ &= \binom{k}{n} (r(t))^n [1 - r(t)]^{k-n}, \quad k \geq n, \end{aligned} \tag{6.10.2}$$

where

$$\begin{aligned} r(t) &= Pr\{\text{a unit arriving in } (0, t), \text{ is still in service at epoch } t\} \\ &= \int_0^{\infty} Pr\{\text{service time of a unit exceeds } (t - x), \\ &\quad \text{given that the unit arrives at epoch } x(<t)\} \\ &\quad \times Pr\{\text{the unit's arrival epoch is } x\} dx. \end{aligned}$$

Now, given that an arrival (from a Poisson process) occurs in $(0, t)$, the interval x in $(0, t)$ is uniformly distributed over $(0, t)$ (see Section 1.5.1). Hence, $r(t)$ equals the probability that a unit's service time does not terminate by time t, given that his arrival is uniformly distributed over the interval $(0, t)$.

Thus,

$$\begin{aligned} r(t) &= \frac{1}{t} \int_0^t Pr\{v \geq (t - x)\} dx \\ &= \frac{1}{t} \int_0^t \{1 - B(t - x)\} dx \\ &= 1 - \frac{1}{t} \int_0^t B(u) du. \end{aligned} \tag{6.10.3}$$

We get

$$\begin{aligned} p_n(t) &= \sum_{k=n}^{\infty} \binom{k}{n} (r(t))^n [1 - r(t)]^{k-n} \frac{e^{-\lambda t}(\lambda t)^{k-n+n}}{k!} \\ &= \sum_{k=n}^{\infty} \frac{e^{-\lambda t}}{n!(k-n)!} \{\lambda t r(t)\}^n \{\lambda t (1 - r(t))\}^{k-n} \\ &= \frac{\{\lambda t r(t)\}^n}{n!} e^{-\lambda t} e^{\lambda t (1 - r(t))}. \end{aligned}$$

Finally,

$$p_n(t) = \frac{\{\lambda t\, r(t)\}^n}{n!} e^{-\{\lambda t r(t)\}}, \qquad n = 0, 1, 2, \ldots. \tag{6.10.4}$$

The distribution of $N(t)$ is Poisson with mean

$$\{\lambda t[r(t)]\} = \lambda \int_o^t \{1 - B(u)\} du. \tag{6.10.5}$$

Thus, $\{N(t), t \geq 0\}$ is a nonhomogeneous Poisson process.

6.10.1.1 Steady-state solution

As $t \to \infty$

$$\lambda t[r(t)] = \lambda \int_0^t \{1 - B(u)\} du \to \frac{\lambda}{\mu},$$

so that the steady-state distribution of the system size (or the number of busy channels) is Poisson with mean λ/μ, irrespective of the magnitude of λ/μ. Thus, when the number of channels is large (so that an arrival always finds an empty channel), the steady-state distribution of the system size is Poisson, irrespective of the form of the service-time distribution and of the magnitude of (λ/μ). It may be recalled that the distribution is truncated Poisson for a c-server (exponential server) $M/M/c$ model.

The busy period distribution of the system $M/G/\infty$ earlier considered by Takács (1962) has been considered by Stadje (1985). Hall (1985) considers heavy traffic approximation.

Notes: $M/G/\infty$: Steady State

(1) Expected busy period T (with at least one server busy):
From Eq. (6.10.4), we have $(a = \lambda/\mu)$

$$p_0 = \lim_{t\to\infty} p_0(t) = e^{-a}, \quad \text{so that}$$

$$\frac{E(T)}{E(T) + E(I)} = 1 - p_0 = 1 - e^{-a}.$$

Thus,

$$E(T) = \frac{1 - p_0}{p_0} = (e^a - 1)/\lambda.$$

(2) The distribution of the number of busy servers is Poisson with mean $a = \lambda/\mu$ and variance a. Since Poisson distribution can be approximated by a normal distribution when a is large (Central limit Theorem), the steady-state

number of busy servers, N_a, is approximately normally distributed with mean a—that is, for a large

$$P\{N_a \geq x\} \cong 1 - \Phi\left(\frac{x-a}{\sqrt{a}}\right),$$

Φ being the DF of standard normal $N(0, 1)$.

(3) It follows from Little's Law that the mean number of steady-state busy servers is always equal to $a = \lambda/\mu$ for all infinite server queues, including the $G/G/\infty$ queue, although the distribution is not Poisson. It is shown that the distribution of busy servers is asymptotically normal for large a even for the $G/G/\infty$ queue.

If $A(t)$ represents the number of arrivals in the interval $(0, t)$ and the squared coefficient of variation of the interval time is

$$c_a^2 = \lim_{t\to\infty} \frac{\text{var}\{A(t)\}}{E\{A(t)\}},$$

and $B(.)$ represents the DF of the service time S. Then for large a,

$$P\{N_a \geq x\} \cong 1 - \Phi\left(\frac{x-a}{\sqrt{za}}\right),$$

where

$$z = 1 + (c_a^2 - 1)\int_0^\infty [1 - B(x)]^2 dx.$$

That is, N_a is asymptotically normal.

When arrival is Poisson $c_a^2 = 1$ and $z = 1$. When S is exponential, $z = (c_a^2 + 1)/2\mu$.

(Borovkov, 1967; Whitt, 1992)

Remark: *$M^X/G/\infty$ in Steady State*

The mean and variance of the number N of busy servers in an $M^X/G/\infty$ queue in steady state are given by

$$E\{N\} = \lambda E(X) E(S),$$
$$\text{var}\{N\} = E(N) + \lambda E\{X(X-1)\}\int_0^\infty [1 - B(x)]^2 dx,$$

where S is the service time with DF $B(.)$ (Tijms, 1994).

Remark: *Invariant or Insensitivity Property.* The steady-state probabilities in the model considered are said to be insensitive of the service-time distribution (the same occurs only through its first moment $1/\mu$). This property is called the *invariant* or *insensitivity property*, and the system is said to be invariant or insensitive. The property is quite a powerful one. It enables one to infer and to derive property of systems with more tractable exponential service-time

distribution. Further, approximations can be made in cases where the conditions for the validity of invariance are *almost* satisfied. Another important aspect is the connection between the invariant property and the product form of state probabilities. The discovery of such a property has contributed, to a large degree, to advancement in applications to computer and communication systems. For a description of the class of queueing models that possess such a property, refer to Dukhovny and Koenigsberg (1981), and also to Disney and König (1985).

$M_t/G/\infty$: Nonhomogeneous $M/G/\infty$ system

Here we consider a nonhomogeneous Poisson process with deterministic arrival rate function

$$\lambda \equiv \{\lambda(t), -\infty < t < \infty\}.$$

The number of arrivals in any interval $[s, t]$ has a Poisson distribution with mean

$$\int_s^t \lambda(u)du.$$

The service times $\{S\}$ are IID with DF $G(.)$ and are independent of the arrival process.

The residual lifetime (S_R) distribution of S has PDF

$$G_R(t) = \frac{1}{E(S)} \int_0^t \{1 - G(u)\}du, \quad t \geq 0$$

and has moments (see Section 1.7.2 and Problem 1.26).

$$E[S_R^k] = \frac{E[S^{k+1}]}{(k+1)E(S)}, \quad k \geq 1.$$

Theorem 6.3. *$Q(t)$, the number of busy servers at time t has a Poisson distribution, and its mean is given by*

$$E\{Q(t)\} = E(S).E\{\lambda(t - S_R)\}.$$

Further, the departure process is a nonhomogeneous Poisson process with time-dependent rate function

$$\delta(t) = E\{\lambda(t - S)\}.$$

For a proof and an interesting discussion on the above result, see Eick et al. (1993) and Foley (1982).

It is assumed that the system $M_t/G/\infty$ started empty at the distant past—that is, at $t = -\infty$.

6.10.2 The model $G/M/c$

We can now generalize the results of Section 6.7.1 for a $G/M/1$ model to the c-server model $G/M/c$.

Define Y_n as the number in the system immediately before the arrival of the nth unit. Then $\{Y_n, n \geq 0\}$ is a Markov chain. We have

$$Y_{n+1} = Y_n + 1 - B_{n+1},$$

where B_{n+1} is the number of units served (or number of departures) during $[T_n = (t_{n+1} - t_n)]$—that is, the interval between the nth and the $(n+1)$th arrivals. B_n is independent of n. However, the service rate during T_n would be state-dependent. We shall now find the transition probabilities

$$p_{ij} = Pr\{Y_n = j \mid Y_{n-1} = i\}.$$

Case (i): $i + 1 < j$. Since j cannot exceed $i + 1$, $p_{ij} = 0$.
Case (ii): $i + 1 \geq j \geq c$. This case implies that all the servers are busy and the rate of service is $c\mu$, so that

$$p_{ij} = g_{i+1-j},$$

where now

$$g_r = \int_o^\infty \frac{e^{-c\mu t}(c\mu t)^r}{r!}\, dA(t), \quad r = 0, 1, 2, \ldots. \tag{6.10.6}$$

Two more cases arise for $c > 1$.
Case (iii): $j \leq i + 1 \leq c$. If an arrival (say, nth arrival) finds i in the system, then as $i \leq c - 1 < c$, he will find an empty channel and will immediately enter service, so that $(i+1)$ customers will be simultaneously receiving service (or so that $(i+1)$ channels will be busy). In order that the next arrival (the $(n+1)$th) finds exactly j on arrival, $(i+1-j)$, customers out of $(i+1)$ receiving service will complete service during the interarrival interval T_n. Thus,

$$p_{ij} = Pr\{\text{out of } (i+1) \text{ customers being served, services of } (i+1-j) \text{ will be completed during an interarrival interval}\}.$$

Conditioning on the length of the interarrival interval and denoting by

$$A \equiv \text{the event that out of } (i+1) \text{ ongoing services, } (i+1-j) \text{ services will be completed in an interval of length } t,$$

we get

$$p_{ij} = \int_0^\infty Pr\{A \mid t\}\, dA(t).$$

Now, $Pr\{\text{one service is completed in an interval of length } t\} = 1 - e^{-\mu t}$. Thus, A is a binomial RV with parameters $n = (i+1)$ and $p = (1 - e^{-\mu t})$. We have

$$Pr\{A \mid t\} = \binom{i+1}{i+1-j}(1 - e^{-\mu t})^{i+1-j}(e^{-\mu t})^{j}, \quad \text{and}$$

$$p_{ij} = \int_0^\infty \binom{i+1}{j}(1 - e^{-\mu t})^{i+1-j}(e^{-\mu t})^{j}\, dA(t). \tag{6.10.7}$$

Case (iv): $i + 1 \geq c > j$. Here the arrival finds all channels busy, and it can enter service only when $(i + 1 - c)$ services are completed and one of the channels becomes free and then the number of busy channels becomes c. Let the duration of this subinterval be H_c and its DF be $H(.), 0 < H_c < (t_{n+1} - t_n)$. In order that the next arrival (arriving at T_{n+1}) finds j, services of $c - j$ must be completed during the remaining duration of the interarrival interval—that is, during the subinterval of length $T - H$.

Now H_c is the sum of $(i + 1 - c)$ IID exponential RVs with parameter $c\mu$ and is thus a gamma variable with parameters $(i + 1 - c)$ and $c\mu$. We have

$$dH(s) = e^{-c\mu s}\frac{(c\mu)(c\mu s)^{i-c}}{(i-c)!}ds. \tag{6.10.8}$$

Conditioning on the interarrival time T and then on the interval H, we get

$$\begin{aligned} p_{ij} &= \int_0^\infty Pr\{(i+1-j) \text{ service completions in time } t\}dA(t) \\ &= \int_0^\infty \int_0^t Pr\{(i+1-j) \text{ service completions in time } t \mid H = s\}dH(s)dA(t) \\ &= \int_0^\infty \int_0^t Pr\{B \mid H = s\}dH(s)dA(t), \end{aligned}$$

where

$$\begin{aligned} Pr\{B \mid H = s\} &= Pr\{(c-j) \text{ service completions out of } c \text{ services in the} \\ &\qquad \text{(sub) interval of length } t - s\} \\ &= \binom{c}{c-j}\left[1 - e^{-\mu(t-s)}\right]^{c-j}\left[e^{-\mu(t-s)}\right]^{j}. \end{aligned}$$

Thus,

$$p_{ij} = \int_0^\infty \left\{ \int_0^t \binom{c}{j}\left[1 - e^{-\mu(t-s)}\right]^{c-j}\left[e^{-\mu(t-s)}\right]^{j} dH(s) \right\} dA(t) \tag{6.10.9}$$

$$= \binom{c}{j}\frac{(c\mu)^{i+1-c}}{(i-c)!}\int_0^\infty \left\{ \int_0^t \binom{c}{j}\left[1 - e^{-\mu(t-s)}\right]^{c-j} e^{-\mu(t-s)j}s^{i-c}e^{-c\mu s}ds \right\} dA(t). \tag{6.10.9a}$$

The limiting arrival point system size probabilities v_j are the unique solutions of

$$\boldsymbol{V} = \boldsymbol{V}\boldsymbol{P}, \qquad \sum v_j = 1.$$

We note that for $i + 1 \geq j \geq c$, we get the same type of expression of p_{ij} (with g_r given by Eq. (6.10.6) for the c-server case in place of g_r given by Eq. (6.9.2) for the single-server case).

Thus, we can proceed in a similar manner as in the single-server case to find v_j for $j \geq c$. We shall get

$$v_j = Cr_o^j, \quad j \geq c, \tag{6.10.10}$$

where C is a constant and r_0 is the root of the equation

$$z - A^*(c\mu - c\mu z) = 0. \tag{6.10.11}$$

The constant C and the first c values, $v_j s$ (i.e., $v_0, v_1, \ldots, v_{c-1}$) are to be determined recursively from the first $(c-1)$ equations of $\boldsymbol{V}\boldsymbol{P} = \boldsymbol{V}$—that is, from

$$v_j = \sum_{i=0}^{\infty} v_i p_{ij}, \quad j = 0, 1, 2, \ldots, c-1$$

and the normalizing relation $\sum_{j=0}^{\infty} v_j = 1$.

From the last relation, we get

$$C = \frac{1 - \sum_{i=0}^{c-1} v_i}{\sum_{j=c}^{\infty} r_0^j} = \frac{1 - \sum_{i=0}^{c-1} v_i}{\frac{r_0^c}{(1-r_0)}}. \tag{6.10.12}$$

6.10.2.1 Waiting-time distribution of a G/M/c queue

The distribution for the queueing time W_Q for a $G/M/c$ queue can be obtained in the same way as for a $G/M/1$ queue. We have

$$\begin{aligned} Pr\{W_Q = 0\} &= \sum_{i=0}^{c-1} v_i = 1 - \frac{Cr_0^c}{(1-r_0)} \quad \text{and} \\ Pr\{W_Q \neq 0\} &= \frac{Cr_0^c}{(1-r_0)}. \end{aligned} \tag{6.10.13}$$

Since waiting-time distribution is a geometric compounding of exponential distributions, the distribution is modified exponential (similar to that of the simple queue). We shall have

$$\begin{aligned} W_Q(t) &= Pr\{W_Q \leq t\} = 1 - Pr\{W_Q \neq 0\} e^{-\mu c(1-r_0)t} \\ &= 1 - \frac{Cr_0^c}{(1-r_0)} e^{-\mu c(1-r_0)t}, \quad t > 0. \end{aligned} \tag{6.10.14}$$

Note: Bulk-arrival multiserver $GI^X/M/c$ queue is considered by Zhao (1994) and the finite buffer model $GI^X/M/c/N$ by Laxmi and Gupta (2000). They also examine batch service model $GI/M^{(b)}/c/N$ (1999).

6.10.3 The model $M/G/c$

This model cannot be analyzed in the same way as the model $G/M/c$. The difficulty arises because of the fact that an embedded Markov chain cannot be extracted from it in the same way as it was done in the case of the $G/M/c$ system. We cannot therefore get any tidy result about the queue-length distribution for this system.

However, Little's formula and its generalized version in terms of moments of higher order hold. As already mentioned in Remarks in Section 2.6, we get, for $r = 1, 2, 3, \ldots,$

$$E\{L^r\} = \frac{\lambda^r E\{W^r\}}{r!} \quad \text{and} \tag{6.10.15}$$

$$E\{L_{(r)}\} = \lambda^r E\{W^r\}, \tag{6.10.16}$$

where

$$L_{(r)} = L(L-1)\cdots(L-r+1).$$

The relation (6.10.16) connects the factorial moments of L with the moments of W. See Brumelle (1972) for derivation of the relations.

6.10.3.1 The loss system $M/G/c/c$

We have discussed the $M/M/c/c$ (Erlang loss system) in Section 3.7 and obtained

$$\begin{aligned} p_k &= Pr\{\text{an arrival finds } k \text{ channels busy}\} \\ &= \frac{(\lambda/\mu)^k/k!}{\sum_{n=0}^{c}(\lambda/\mu)^n/n!}, \quad 0 \le k \le c. \end{aligned}$$

Attempts have been made since Erlang's time to generalize the preceding formula (for example, by Palm, Kosten, Pollaczek, Vaulot, Sevast'yanov, and Fortet). It has been found that the formula holds for arbitrary distribution of the service time—that is, for $M/G/c/c$. (It was known to Erlang also.) The first rigorous proof of the validity of the formula was given by Sevast'yanov (1956). Fortet (1956) obtained the result under the assumption that the distribution of service time is absolutely continuous. A proof based on the basic ideas of Sevast'yanov's proof (in Russian) is given in Gnedenko and Kovalenko (1968). A brief outline of the proof is given next.

If ($N(t)$ is the queue length at epoch t (here the queue length equals the number of busy channels), then the process $\{N(t), t \geq 0\}$ is non-Markovian. Sevast'yanov uses the supplementary variable technique that consists of inclusion of additional variables.

Let us assume that at some epoch t, $N(t)$, the number of the busy channels equals $k, 0 \leq k \leq c$ and $N(t-0) \neq k$. Let the servers busy at epoch t be given serial numbers $1, 2, \ldots, k$ in random order and let $f_i(t)$ denote the time elapsed from epoch t until the server with serial number i completes the service. ($f_i(t)$ is the remaining service time from epoch t of server number i.) The vector stochastic process

$$\boldsymbol{V}(t) = \{N(t);\ f_1(t), \ldots, f_k(t)\} \tag{6.10.17}$$

is a Markov process. Denote

$$F_k(t; x_1, \ldots, x_k) = Pr\{N(t) = k;\ f_1(t) \leq x_1, \ldots, f_k(t) \leq x_k\}. \tag{6.10.18}$$

Then

$$F_k(t; \infty, \ldots, \infty) = Pr\{N(t) = k\} = p_k(t) \quad \text{and}$$
$$p_k = \lim_{t \to \infty} F_k(t; \infty, \ldots, \infty). \tag{6.10.19}$$

Denote

$$F_k(x_1, \ldots, x_k) = \lim_{t \to \infty} F_k(t; x_1, \ldots, x_k). \tag{6.10.20}$$

The Chapman-Kolmogorov equations are then obtained. It is shown that

$$F_k(x_1, \ldots, x_k) = \frac{\lambda^k}{k!} F_o \prod_{i=1}^{k} \int_0^{x_i} [1 - B(u)]du \tag{6.10.21}$$

satisfies the equations $B(.)$ being the service time DF. We get

$$p_k = \lim_{x_i \to \infty} F(x_1, \ldots, x_k) = \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} F_0, \quad 0 \leq k \leq c.$$

Using the normalizing condition, one gets Erlang's formula. For details of Sevast'yanov's proof, refer to Gnedenko and Kovalenko (1968). Also see Takács's (1969) paper for a general review.

Notes:

1. We have here a model having an invariant (or insensitivity) property discussed in the remark to Section 6.10.1.
2. The insensitivety property is no longer true in the case of batch arrivals.

Remarks:

(1) The Erlang loss and delay formulas have been a subject of continued interest, especially problems of asymptotic analysis, approximations, inequalities, bounds, convexity of performance measures (of interest in the study of optimization models), and so on. These problems have been studied, among others, by Jagerman (1974), Sobel (1980), Akimaru and Takahashi (1981), Harel (1987), Harel and Zipkin (1987), and Krishnan (1990).

See also Section 3.7.1 for properties of the Erlang Loss Formula (discussed for the $M/M/c/c$ model, which also hold for the $M/G/c/c$ model).

Kaufmann (1979) discusses the busy probability of loss system.

(2) *$M^X/G/c/c$ Loss Model*

Tijms and Hogenkamp (1995) propose an approximation of the loss probability P (long-run fraction of customers who are rejected) for the above model (steady state).

Let N be the number of busy servers in an $M^X/G/\infty$ queue in steady state and let

$$p = \frac{E(N)}{\text{var}(N)} \quad \text{and} \quad r = \frac{pE(N)}{1-p}$$

(see Tijms (1994) for $E(N)$ and $\text{var}(N)$).

For $X \neq 1$, $p < 1$, Let

$$p_j = \binom{r+j-1}{j} p^r (1-p)^j, \quad j = 0, 1, 2, \ldots,$$

and

$$p_j^* = \frac{p_j}{\sum_{k=0}^{c} p_k}, \quad j = 0, 1, 2, \ldots, c.$$

Then P can be approximated by

$$P \simeq \frac{1}{E(X)} \sum_{j=0}^{c} p_j^* \sum_{k=c-j+1}^{\infty} (j+k-c) Pr(X = k).$$

The approximation is exact for the $M^X/M/c/c$ model and for X having geometric distribution. See Tijms and Hogenkamp (1995).

6.11 Queues with Markovian Arrival Process

Before concluding this chapter, here is a brief description of a more general arrival process that includes the Poisson process as a special case. This class of tractable Markovian Arrival Processes (MAP) was introduced by Neuts (1979) and has been studied extensively by him and other researchers after him. These have been used in modeling serval situations arising in communication systems.

Consider a continuous-time Markov process with exactly *one* state, which is visited successively. The sojourn interval is exponential—say, with parameter λ_i. At the completion of one sojourn time, the Markov process visits the same state, and another sojourn time (having an independent exponential distribution with the same parameter λ) starts. If an arrival is associated with the revisit to the state, then the number of arrivals in an interval of length t corresponds to the occurrence of a Poisson process with parameter λt.

This situation is generalized as follows. Suppose that a Markov process with state space $\{1, 2, \ldots, m+1\}$ has m transient states $\{1, 2, \ldots, m\}$ and one absorbing state $m+1$, such that absorption starting from any transient state to the absorbing state is certain. Suppose that the sojourn time at the transient state i is exponential with parameter λ_i. At the completion of sojourn at state i, there are two possibilities: The Markov process enters the absorbing state $(m+1)$ and immediately reenters another transient state j with probability p_{ij}, or it enters another transient state $j, 1 \leq j \leq m, j \neq i$ with probability q_{ij}. We have then

$$\sum_{j=1}^{m} p_{ij} + \sum_{\substack{j=1 \\ j \neq i}}^{m} q_{ij} = 1, \quad 1 \leq i \leq m. \tag{6.11.1}$$

Define

$$\begin{aligned} D_{ij} &= \lambda_i p_{ij}, \quad 1 \leq i \leq m \\ C_{ij} &= \lambda_i q_{ij}, \quad 1 \leq i, j \leq m \\ \text{and} \quad C_{ii} &= -\lambda_i. \end{aligned}$$

Associate an entry into the absorbing state with an arrival. Then the probability that in an infinitesimal interval of length dt, the process enters the absorbing state $(m+1)$ (and then an arrival is said to occur), given that it was in state i, is $D_{ij}\, dt$, and the probability that the process enters another state j, given that it was in state i, is $C_{ij}\, dt$, $j \neq 1$.

Denote the matrix (C_{ij}) by $\boldsymbol{C}$ and the matrix (D_{ij}) by $\boldsymbol{D}$. Then $\boldsymbol{C} + \boldsymbol{D}$ is the infinitesimal generator of the underlying Markov process describing the transitions among the transient states $\{1, 2, \ldots, m\}$.

We thus have a generalization of a Poisson process and get a ***M**arkovian **A**rrival **P**rocess* (***MAP***). For $m = 1$, $D = \lambda$, and $C = -\lambda$, we get a Poisson process. An $MAP/G/1$ queue is a matrix generalization of an $M/G/1$ queue.

A MAP process is a special case of the general Semi-Markov Process (SMP).

By introducing arrivals in batches with IID batch sizes (where each arrival corresponds to a batch arrival in an MAP), one gets a ***B**atch **M**arkovian **A**rrival **P**rocess* (***BMAP***). $BMAP/G/1$ queue is a matrix formalism of an $M^X/G/1$ queue. A BMAP process allows inclusion of dependent interarrival times, non-exponential interarrival times, correlated batch sizes, and so on. Such a process

includes both renewal type (including Neuts's PH renewal process) and non-renewal processes such as the Markov Modulated Poisson Process (MMPP).

Queueing systems with such processes have been discussed extensively in the literature. References may be made to Neuts (1979), Ramaswami (1980), Lucantoni *et al.* (1990), Lucantoni (1993), Latouche and Ramaswami (1999), and the references therein. Transient state behavior has been discussed in Lucantoni *et al.* (1994). Discrete-time systems have also been studied.

Problems and Complements

6.1. Consider an $M/G/1$ queueing system in steady state. Show that $\{k_j\}$ is geometric *iff* $\{p_j\}$, $j = 0, 1, 2, \ldots$, is geometric. The result also holds for the zero-truncated geometric, that is, for $j = 1, 2, \ldots$ (Rego and Szpankowski, 1989).

6.2. Consider an $M/G/1$ system in steady state having service-time distribution $B(.)$ with finite first- and second-order moments. Show that any one of the following statements implies the other two:

(i) $B(.)$ is exponential.

(ii) $\{k_j\}$, $j = 0, 1, 2, \ldots$, is geometric with mass function

$$k_j = ab^j, \quad a = 1 - b = \frac{\mu}{(\lambda + \mu)}.$$

(iii) $\{p_j\}$, $j = 0, 1, 2, \ldots$, is geometric with mass function

$$p_j = \left(\frac{2a - 1}{a}\right)\left(\frac{1 - a}{a}\right)^j.$$

(Rego and Szpankowski, 1989)

6.3. Suppose that the busy period T is initiated by $m(\geq 1)$ customers (i.e., there are m customers at the commencement of the busy period). Then the PGF of the number $N_m(T)$ served during the busy period is given by $[P(z)]^m$, where $P(z)$ is the PGF of the number served during the busy period initiated by a single customer.
Show that, for $n = m, m + 1, \ldots$

(i)

$$Pr\{N_m(T) = n\} = \frac{m}{n}\binom{2n - m - 1}{n - 1}\frac{\rho^{n-m}}{(1 + \rho)^{2n-m}}$$

in the case of an $M/M/1$ queue (see Section 3.9.4.2);

(ii)

$$Pr\{N_m(T) = n\} = \frac{m}{(n-m)!} n^{n-m-1} \rho^{n-m} e^{-n\rho}$$

in the case of an $M/D/1$ queue. This is called the Borel-Tanner distribution. Find the mean and variance of $N_m(T)$.

6.4. $M/G/1$ queue-length distribution (Willmott, 1988)

(a) Refer to Section 6.3 and the Pollaczek-Khinchin formula (6.3.7). Verify that $K(s)$ is a PGF; so also is

$$G(s) = \sum g_n s^n = \frac{K(s) - 1}{\rho(s-1)}, \tag{A}$$

where g_n is given by $(1 - \sum_{j=0}^{n} k_j)/\rho, n = 0, 1, 2, \ldots$. Show that Eq. (6.3.7) can be decomposed into

$$V(s) = Q(s)K(s), \tag{B}$$

where

$$Q(s) = \sum q_n s^n = \frac{1-\rho}{1-\rho G(s)} \tag{C}$$

is the PGF of a compound geometric distribution. Hence, show that

$$v_n = \sum_{j=0}^{\infty} q_j k_{n-j}, \quad n = 0, 1, 2, \ldots, \tag{D}$$

and that q_n can be expressed as

$$q_n = \sum_{m=0}^{\infty} (1-\rho)\rho^m g_n^{(m)*}, \tag{E}$$

where $g_n^{(m)*}$ is the coefficient of s^n in $[G(s)]^m$. Show that $Q(s)$ can also be expressed as

$$Q(s) = \frac{1-\rho_1}{1-\rho_1 G_1(s)}, \tag{F}$$

where

$$\begin{aligned} \rho_1 &= \frac{\rho(1-g_0)}{(1-\rho g_0)} \quad \text{and} \\ G_1(s) &= \frac{G(s) - g_0}{1 - g_0}, \end{aligned} \tag{G}$$

whence

$$g_n = \sum_{m=0}^{n} (1-\rho_1)\rho_1^m g_1^{(m)*}, \quad n = 0, 1, 2, \ldots \tag{H}$$

(expressed as a finite sum as an alternate to (E)) where $g_1^{(m)*}$ = coefficient of s^n in $[G_1(s)]^m$.

(b) Obtain as a particular case the corresponding results for an $M/M/1$ queue.

(c) How would you generalize the results of (a) to bulk-arrival queues?

6.5. Show that the steady-state probabilities $p_j(=v_j)$, $j = 0, 1, 2, \ldots$ in an $M/G/1$ queue satisfy the recursive relation

$$p_j = \lambda a_{j-1} p_0 + \lambda \sum_{i=1}^{j} a_{j-k} p_k, \quad j = 1, 2, \ldots, \tag{A}$$

where

$$a_n = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} \{1 - B(t)\} dt, \quad n = 0, 1, 2, \ldots \quad \text{and}$$

$B(t)$ is service-time DF.

Deduce that

$$\begin{aligned} P(s) &= \sum_{j-0}^{\infty} p_j s^j, \quad |s| \le 1 \\ &= (1-\rho)\frac{1 - \lambda(1-s)A(s)}{1 - \lambda(1-s)}, \end{aligned} \tag{B}$$

where

$$\begin{aligned} A(s) &= \sum_{n=0}^{\infty} a_n s^n \\ &= \int_0^\infty [1 - B(t)] e^{-\lambda(1-s)t}\, dt. \end{aligned}$$

(See Tijms (1994).)

6.6. Higher moments of W and W_q in an $M/G/1$ queue: Recursive formulas for computation.

Takagi and Sakamaki (1995) discuss symbolic moment calculation. Obtain the following recursive formulas:

(a) $$E[W_q^n] = \frac{\lambda}{1-\rho} \sum_{m=1}^{n} \binom{n}{m} \frac{b^{(m+1)}}{m+1} E[W_q^{n-m}] \quad 1 \le m \le n-1,$$

$$n = 1, 2, \ldots,$$

where

$$b^{(m+1)} = (m+1)th \quad \text{moment of service time distribution;}$$

(b) $$E[W^n] = \sum_{m=0}^{n} \binom{n}{m} b^{(m)} \, E[W^{n-m}] \quad n = 0, 1, 2, \ldots .$$

Also show that

(c) $$E[W^n] = E\left[W_q^n\right] + \frac{n}{\lambda} E\left[W_q^{n-1}\right] \quad n = 2, 3, \ldots .$$

Hints: Write Eq. (6.3.11a) [Eq. (6.3.10)] in a form putting only $(1 - \rho)$ on the RHS; then differentiate both sides n times and put $s = 0$.

6.7. Consider an $M/G/1$ system with X_n as defined in Eq. (6.3.1) an $K(.)$ as given in Eq. (6.2.2). Let $\mu(n)$ be defined by

$$\mu(n) = \begin{cases} \inf\{m \geq 0: \ X_{n+m} = 0\}, & \text{if this set is nonempty;} \\ \infty & \text{otherwise,} \end{cases}$$

$$n = 0, 1, 2, \ldots$$

Show that for $0 \leq z \leq 1$, $n = 0, 1, 2, \ldots$

$$E\left[Z^{X_n}\right] = E\left\{\left[\frac{z}{K(z)}\right]^{\mu(n)}\right\}.$$

(Baccelli and Makowski, 1989)

6.8. Consider an $M/G/1$ having an exceptional service for the first unit in a busy period (this may be necessitated for doing some extra work or for making some initial work to start services before serving the first customer immediately after the idle period is over) (Takagi, 1991).

Denote by $B_0(.)$, $B_0^*(.)$, b_0, $b_0^{(2)}$, the DF, the LST, the first and the second moments, respectively, of the exceptional first service, and by $B(.)$, $B^*(.)$, b, $b^{(2)}$ these corresponding to the other service times.

Denote

B = service time of a unit in general
= B_0 or B according as the unit is the first to be served in a busy period or later in the busy period

B_{1R} = residual general service time

λ = arrival rate, $a = \lambda b$

I = idle period

Show that

(a) the fraction of time the server is busy is given by

$$\rho = \frac{\lambda b_0}{1 - a + \lambda b_0}.$$

(b) the expected busy period is given by

$$E(T) = \frac{b_0}{1 - a}$$

and that it satisfies

$$E(T) = \frac{\rho}{1 - \rho} E(I).$$

(c) the expected queueing time is

$$W_Q = \frac{\rho}{1 - a} E(B_{1R}) = \frac{\lambda b^{(2)}}{2(1 - a)} + \frac{\lambda\left(b_0^{(2)} - b^{(2)}\right)}{2(1 - a + \lambda b_0)}.$$

Denote by $W^*(s)$ the LST of the queueing time of this model and by $W^*(s \mid \text{system busy})$ LST of the conditional waiting time given that the system is busy.
Show that

$$W^*(s \mid \text{system is busy}) = \frac{(1 - a)(1 - B_0^*(s))}{b_0(s - \lambda + \lambda B^*(s)).}$$

Show that this can be factorized into two factors A and B, where

$A \equiv$ the LST of the waiting time of the standard $M/G/1$ queue
$B \equiv$ the LST of the residual first service time B_0

How do you interpret this result?
Show also that

$$W^*(s) = \left(\frac{1 - a}{1 - a + \lambda b_0}\right)\left(\frac{s - \lambda B_0^*(s) + \lambda B^*(s)}{s - \lambda + \lambda B^*(s)}\right).$$

Note that when $B_0(.) \equiv B(s)$, then one gets the corresponding results for the standard $M/G/1$ queue.

6.9. Show that (with notations as in Section 6.4.4)

$$E\left(T_b^2\right) = \frac{\lambda E(v^2)}{(1 - \rho)^3} E(T_0) + \frac{\rho^2}{(1 - \rho)^2} E\left(T_0^2\right)$$

$$E\left(T_c^2\right) = \frac{\lambda E(v^2)}{(1 - \rho)^3} E(T_0) + \frac{E\left(T_0^2\right)}{(1 - \rho)^2}.$$

(Miller, 1975)

6.10. (a) Consider an $M/G/1/n$ finite queue with total space capacity n. Let T_0 be the delay with DF $H(t)$ and LST $H^*(s)$, let v be the service time with DF $B(t)$ and LST $B^*(s)$, and let $A_n(s)$ be the LST of ordinary busy period T and $A_n^d(s)$ be the LST of delay busy period T_d. Denote

$$u_k(s) = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-(\lambda+s)t}\, dB(t)$$

$$v_k(s) = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-(\lambda+s)t} dH(t).$$

Show that

$$A_n(s) = \frac{u_0(s)}{\left[1-\sum_{k=1}^{n-1} u_k(s)\prod_{j=n-k+1}^{n-1} A_j(s) - \sum_{k=n}^{\infty} u_k(s)\prod_{j=1}^{n-1} A_j(s)\right]}$$

and

$$A_n^d(s) = v_0(s) + \sum_{k=1}^{n-1} v_k(s) \prod_{j=n-k+1}^{n} A_j(s) + \sum_{k=n}^{\infty} v_k(s) \prod_{j=1}^{n} a_j(s).$$

For an ordinary busy period T, writing

$$p_k = u_k(0), \qquad P_j = 1 - \sum_{k=0}^{j} p_k, \quad \text{and}$$

$$b_n = \text{mean busy period of } M/G/1/n,$$

one gets

$$b_1 = \frac{E(v)}{p_0}$$

$$b_n = \frac{\left[b_{n-1} - \sum_{j=1}^{n-1} b_j p_{n-j}\right]}{p_0}, \quad n \geq 2.$$

Assume that $\rho < 1$, then $\lim_{n\to\infty} b_n$ exists. The generating function $B(z)$ of $\{b_n\}$ is given by

$$B(z) = \sum b_n z^n = \frac{z^{E(v)}}{B^*(\lambda - \lambda z) - z} \quad \text{and}$$

$$E(T) = \lim_{n\to\infty} b_n = \lim_{z\to 1}(1-z)\,B(z).$$

(Miller, 1975; see also Harris, 1971)

(b) For an $M/M/1/n$ queue, deduce, that

$$b_n = \frac{1}{\mu}\frac{1-\rho^{n+1}}{1-\rho}, \quad \lambda \neq \mu$$
$$= \frac{n+1}{\mu}, \quad \lambda = \mu.$$

6.11. Losses in an $M/G/1/n$ queue with finite buffer.
Righter (1999) discusses this problem, earlier considered by Abramov (1997).

Let I_n be the indicator variable that is 1 if there is a loss during a busy period, let $p_n = E\{I_n\}$, and let L_n be the number of loss (during a busy period) in an $M/G/1/n$ queue.

(a) Show that

$$E\{L_{n+1}\} = \frac{E\{L_n\} - p_n}{1 - p_n}, \quad n \geq 1.$$

In particular, if $\lambda = \mu$, then

$E\{L_1\}$ = expected number of arrivals during a service time of a unit = 1

so that $E\{L_n\} = 1$ for all $n \geq 1$ and further that $\text{var}\{L_n\}$ increases with n.

(b) Show that, for all $n \geq 1$
(i) if $\lambda < \mu$, then $E\{L_n\} < 1$ and $E\{L_n\}$ is decreasing in n, and
(ii) if $\lambda > \mu$, then $E\{L_n\} > 1$ and $E\{L_n\}$ is increasing in n.

(c) Show that for an $M^X/G/1/n$ queue, $\lambda E(X) = \mu$, and for all $n \geq 1$

$$E\{L_{n+1}\} \leq \frac{E\{L_n\} - p_n}{1 - p_n}, \qquad E\{L_1\} = 1,$$

and $E\{L_n\}$ is decreasing in n.

(See Righter (1999).)

6.12. (a) Consider a $G/G/1$ system. Denote by R the arrival-epoch system size. Denote by N the general-time-system size. Denote for $n = 0, 1, 2, 3, \ldots$.

$a_n = v_n = Pr(R = n)$
$\hat{v}_n = Pr(R > n)$
$p_n = Pr(N = n)$,
$\hat{p}_n = Pr(N > n)$, and
V = the remaining service time of the service in progress provided the server is occupied
$b_n = E(V \mid R = n), \quad n = 1, 2, \ldots$.

Let λ and μ be the arrival and service rates, respectively, and let

$$\begin{aligned} A_n(x) &= \text{DF of } n\text{th interarrival interval} \\ B(x) &= \text{DF of service time} \\ H(x) &= \text{DF of actual waiting time.} \end{aligned}$$

Then show that the following relations exist between the queue size and actual waiting time:

$$v_n = \int_0^\infty A_{n+1}(x)\,dH(x)$$

$$\hat{p}_n = \rho \int_0^\infty A_n(x)\,dH^* B_1(x),$$

where

$$B_1(x) = \mu \int_0^\infty [1 - B(u)]\,du$$

and * denotes convolution (Mori, 1980).

(b) Show that the following relations hold good:

$$p_n = \lambda\left(b_n v_n + \frac{v_n}{\mu}\right), \quad n = 1, 2, \ldots, \quad \text{and}$$

$$b_n = \frac{(\hat{p}_n - \rho \hat{v}_n)}{\lambda v_n}, \quad n = 1, 2, \ldots$$

Deduce the Pollaczek-Khinchin mean value formula for $M/G/1$. Also deduce the relation $p_n = pv_{n-1}, n = 1, 2, \ldots$, for the $G/M/1$ system (Fakinos, 1982).

6.13. Busy Period of an $M/G/\infty$ Queue

Define a busy period of an infinite-server queue with Poisson input as the interval during which at least one customer is present and is receiving service. Denote

$$\begin{aligned} B(.) &= \text{DF of service-time distribution} \\ C(t) &= 1 - \exp\left\{-\lambda \int_0^t [1 - B(x)]\,dx\right\}, \quad t > 0 \\ c(t) &= C'(t) = \lambda[1 - B(t)][1 - C(t)] \\ H(t) &= \text{DF of busy period distribution} \end{aligned}$$

Show that

$$H(t) = 1 - \frac{1}{\lambda} \sum_{n=1}^{\infty} c^{(n)*}(t),$$

where $c^{(n)*}(t)$ is the n-fold convolution of $c(t)$ with itself (Stadje, 1985).

Show that the expected duration of the busy period is $(e^{\lambda/\mu} - 1)/\lambda$.

6.14. $M/G/1$ queue: interdeparture interval

Let γ be the interdeparture interval of an $M/G/1$ queue in steady state. Show that the LST of γ is given by

$$D^*(s) = \rho B^*(s) + (1-\rho)\{A^*(s)B^*(s)\},$$

where $A^*(s)$, $B^*(s)$ are the LST of interarrival and service time, respectively. Hence find $E(\gamma)$ and $\text{var}(\gamma)$.

Show that the DF of γ for an $M/D/1$ queue is given by

$$\begin{aligned} D(t) &= 0, \quad t < T \\ &= 1 - (1-\rho)e^{-\lambda(t-T)}, \quad t \geq T, \end{aligned}$$

where T gives the constant duration of the service time.

References and Further Reading

Abramov, V. M. (1997). On a property of refusal stream. *J. Appl. Prob.* **34**, 800–805.

Akimaru, H., and Takahashi, H. (1981). Asymptotic expansion for Erlang loss function and its derivative. *IEEE Trans. Comm.* **29**, 1257–1260.

Asmussen, S. (1987). *Applied Probability and Queues,* Wiley, New York.

Baccelli, F., and Makowski, A. M. (1986). Martingale arguments for stability: the $M/GI/1$ case. *Systems Control Lett.* **6**, 181–186.

Baccelli, F., and Makowski, A. M. (1989). Dynamic, transient and stationary behavior of the $M/GI/1$ queue via martingales. *Annals Prob.* **17**, 1691–1699.

Bertsimas, D., and Papaconstantinou, X. (1988). On the steady-state solution of the $M/C_2(a,b)/s$ queueing system. *Transportation Sci.* **22**, 125–138.

Bhat, U. N., and Basawa, I. (1992). *Queueing and Related Models,* Clarendon Press, Oxford.

Borovkov, A. A. (1967). On limit laws for service processes in multi-channel systems. *Siberian Math. J.* **8**, 746–763.

Borovkov, A. A. (1984). *Asymptotic Methods in Queueing Theory,* John Wiley, New York.

Borthakur, A. (1975). On busy period of a bulk queueing system with a general rule for bulk service. *Opsearch* **12**, 40–46.

Borthakur, A., and Medhi, J. (1974). A queueing system with arrival and service in batches of variable size. *Cah. du. Centre d'Et. de Rech Oper.* **16**, 117–126.

Borthakur, A., Medhi, J., and Gohain, R. (1987). Poisson input queueing system with startup time and under control operating policy. *Comp. & Opns. Res.* **14**, 33–40.

Boxma, O. J., and Yechiali, U. (1997). An $M/G/1$ queue with multiple types of feedback and gated vacations. *J. Appl. Prob.* **34**, 773–784.

Brandt, A. (1987). On stationary queue length distributions for $G/M/s/r$ queues. *Queueing Systems* **2**, 321–322.

Brumelle, S. L. (1972). A generalization of $L = \lambda W$ to moments of queue lengths and waiting times. *Opns. Res.* **20**, 1127–1136.

Burke, P. G. (1975). Delays in single server queues with batch input. *Opns. Res.* **23**, 830–833.

Chae, K. C., and Lee, H. W. (1995). $M^X/G/1$ vacation models with N-policy: heuristic interpretation of the mean waiting time. *J. Opnl. Res. Soc.* **46**, 258–264.

Chaudhry, M. L., and Gupta, U. C. (1999). Modeling and analysis of $M/G(a,b)/1/N$ queue—a simple alternative approach. *Queueing Systems* **31**, 95–100.

Chaudhry, M. L., and Templeton, J. G. C. (1983). *A First Course in Bulk Queues,* Wiley, New York.

Chaudhry, M. L., Templeton, J. G. C., and Medhi, J. (1992). Computational results of multiserver bulk-arrival queues with constant service time $M^X/D/c$. *Opns. Res. Suppl.* **40**, S229–S238.

Cohen, J. W. (1982). *The Single Server Queue*, 2nd ed., North-Holland, Amsterdam.

Conway, R. W., Maxwell, W. L., and Miller, L. W. (1967). *Theory of Scheduling*, Addison-Wesley, Reading, MA.

Cooper, R. B. (1981). *Introduction to Queueing Theory*, Edward Arnold, London.

Cooper, R. B. (1990). Queueing Theory in *Handbooks of Operations Research and Management Science* (Eds. D. P. Heyman and M. J. Sobel), Vol. 2, 469–518, North Holland, Amsterdam.

Cosmetatos, G. P. (1976). Some approximate equilibrium results for the multiserver queue $M/G/r$. *Opnl. Res. Qrly* **27**, 615–620.

Cox, D. R. (1955). The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc. Camb. Phil. Soc.* **51**, 433–441.

Disney, R. L., Farell, R. L., and de Morais, P. R. (1973). A characterization of $M/G/1$ queues with renewal departure processes. *Mgmt. Sci.* **19**, 1222–1228.

Disney, R. L., and König, D. (1985). Queueing networks: a survey of their random processes. *SIAM Review* **27**, 335–403.

Dukhovny, I. M., and Koenigsberg, E. (1981). Invariance properties of queueing networks and their applications to computer communication systems. *INFOR* **19**, 185–204.

Eick, S. G., Massey, W. A., and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Opns. Res.* **41**, 731–742.

Enns, E. G. (1969). The trivariate distribution of the maximum queue length, the number of customers served and the duration of the busy period for the $M/G/1$ queueing system. *J. Appl. Prob.* **6**, 154–161.

Fabens, A. T. (1961). The solution of queueing and inventory models by semi-Markov processes. *J. Roy. Stat. Soc.* **B23**, 113–127.

Fabens, A. T., and Perera, A. G. A. D. (1963). A correction to Fabens (1961). *J.R.S.S.* **B 25**, 455–456.

Fakinos, D. (1982). The expected remaining service time in a single server queue. *Opns. Res.* **30**, 1014–1017.

Feller, W. (1968; 1966). *An Introduction to Probability Theory and Its Application*, Vol. 1, 3rd ed.; Vol. II, Wiley, New York.

Fischer, M. J. (1974). The waiting time in the $E_k/M/1$ queueing system. *Opns. Res.* **22**, 898–902.

Foley, R. D. (1982). The non-homogeneous $M/G/\infty$ queue, *Opsearch* **19**, 40–48.

Fortet, R. (1948). Sur la probabilité de perte d'un appel téléphonique, *C. R. Acad. Sc. Paris* **226**, 1502–1504.

Fortet, R. (1956). Random distributions with an application to telephone engineering. *Proc. Berk. Sym. Math. Stat. & Prob.* **2**, 81–88.

Fujiki, M., and Gambe, E. (1980). *Teletraffic Theory* (in Japanese), Maruzen, Tokyo.

Gebhardt, Von D. (1973). Die Ermittlung von Kenngrössen für das Wartesystem $M/G/I$ mit beschränktem Watreraum. *Zeit. fur Opns. Res.* **17**, 207–216.

Glynn, P. W., and Whitt, W. (1991). A new view of the heavy traffic limit theorems for infinite server queues. *Adv. Appl. Prob.* **23**, 188–209.

Gnendenko, B. V., and Kovalenko, I. N. (1968). *Introduction to Queueing Theory* (2nd ed., 1989, Birkhäusser). Israel Prog. for Sci. Tran. Jerusalem.

Gold, H., and Tran-Gia, P. (1993). Performance analysis of batch service queue arising out of manufacturing and system modeling. *Queueing Systems* **14**, 413–426.

Gross, D., and Harris. C. M. (1985). *Fundamentals of Queueing Theory*, 2nd ed. (3rd ed. 1998) Wiley, New York.

Haight, F. A. (1961). A distribution analogous to Borel-Tanner. *Biometrika* **48**, 167–173.

Hajek, B. (1983). The proof of a folk theorem on queueing delay with applications to routing in networks. *J. Ass. Comp. Mach.* **30**, 834–851.

Hall, P. (1985). Heavy traffic approximations for busy period in an $M/G/\infty$ queue. *Stoch. Process & Appl.* **19**, 259–269.

Harel, A. (1987). Sharp bounds and simple approximations for the Erlang delay and loss formulas. *Mgmt. Sci.*

Harel, A., and Zipkin, P. (1987). Strong convexity results for queueing systems. *Opns. Res.* **35**, 405–418.

Harris, T. J. (1971). The remaining busy period of a finite queue. *Opns. Res.* **19**, 219–223.

Henderson, W. (1972). Alternative approaches to the analysis of $M/G/1$ and $G/M/1$ queue. *J. Oper. Res. Soc. Japan* **15**, 92–101.

Heyman, D. P. (1968), Optimal control policies for $M/G/1$ queueing systems. *Opns. Res.* **16**, 362–382.

Humblet, P. A. (1982). Determinism minimizes waiting time in queues. Technical Report, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

Jacob, M. J., Krishnamoorthy, A. K., and Madhu-Soodanan, T. P. (1988). Transient solution for a finite capacity $M/G(1, b)/1$ queue. *Naval Res. Log.* **35**, 437–441.

Jagerman, D. L. (1974). Some properties of the Erlang loss function. *Bell. Syst. Tech. J.* **53**, 525–551.

Kaufman, J. S. (1979). The busy probability in $M/G/N/N$ loss systems. *Opns. Res.* **27**, 204–206.

Keilson, J. (1965). *Green's Function Methods in Probability Theory,* Charles Griffin, London.

Keilson, J., and Kooharian, A. (1960). Time dependent queueing processes. *Ann. Math. Stat.* **31**, 104–112.

Kendall, D. G. (1951). Some problems in the theory of queues. *J. Roy. Stat. Soc.* **B13**, 151–185.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of Markov chains. *Ann. Math. Stat.* **24**, 338–354.

Kingman, J. F. C. (1963). Poisson counts for random sequence of events. *Ann. Math. Stat.* **34**, 1217–1232.

Kleinrock, L. (1975). *Queueing Systems,* Vol. 1, Wiley, New York.

Krishnan, K. R. (1990). The convexity of loss rate in an Erlang loss system and sojourn in a delay system with respect to arrival and service rates, *IEEE Trans on Comm.* **38.**

Latouche, G., and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling,* Am. Stat. Ass. & Soc. Ind. & Appl Math. (SIAM), Philadephia, PA.

Laxmi, P. V., and Gupta, U. C. (1999). On the finite buffer bulk service queue with general independent arrivals: $GI/M^{[b]}/1/N$. *Opns. Res. Lett.* **25**, 241–245.

Laxmi, P. V., and Gupta, U. C. (2000). Analysis of finite-buffer multiserver queue with group arrivals: $GI^X/M/c/N$. *Queueing Syst.* **36**, 125–140.

Lemoine, A. J. (1976). On random walks and stable $GI/G/1$ queues. *Maths. Opns. Res.* **1**, 159–164.

Louchard, G., and Latouche, G. (Eds.) (1983). *Probability Theory and Computer Science,* Part II: *Queueing Models.* Academic Press, New York.

Lu, F. V., and Serfozo, R. F. (1984). Queueing decision processes with monotone hysteric optimal policies. *Opns. Res.* **32**, 1116–1132.

Lucantoni, D. M. (1993). The *BMAP/G/1 Queue: A Tutorial in Models and Technique for Performance Evaluation of Computer and Communication Systems,* Springer, New York.

Lucantoni, D. M., Choudhury, G. L., and Whitt, W. (1994). The transient *BMAP/G/1* queue. *Comm. Statist. Stoch. Models* **10(1)**, 145–182.

Lucantoni, D. M., Meier-Hellstern, K. S., and Neuts, M. F. (1990). A single-server queue with server vacations and a class of non-renewal arrival processes. *Adv. Appl. Prob.* **22**, 676–705.

Madan, K. C. (2000). An $M/G/1$ queue with second optional service. *Queueing Systems* **34**, 37–46.

Medhi, J. (1984). *Recent Developments in Bulk Queueing Models,* Wiley Eastern, New Delhi, India.

Medhi, J. (2001). Response Time in an $M/G/1$ Queueing System with Bernoulli Feedback in *Recent Developments in Operational Research* (Eds. M. L. Agarwal and K. Sen), pp. 249–259, Narosa, New Delhi.

Medhi, J. (2002). A single server Poisson input queue with a second optional channel. *Queueing Systems*, **42**, 239–242.

Medhi, J., and Templeton, J. G. C. (1992). A Poisson input queue under N-policy and with general start-up time. *Comps. & Opns. Res.* **19**, 35–41.

Miller, L. (1975). A note on the busy period of an $M/G/1$ finite queue. *Opns. Res.* **23**, 1179–1182.

Mohanty, S. G. (1972). On queues involving batches. *J. Appl. Prob.* **9**, 430–435.

Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, Academic Press, New York.

Mori, M. (1980). Relation between queue size and waiting time distributions. *J. Appl. Prob.* **17**, 822–830.

Neuts, M. F. (1966). An alternative proof of a theorem of Takács on $GI/M/I$ queue. *Opns. Res.* **14**, 313–317.

Neuts, M. F. (1967). A general class of bulk queues with Poisson input. *Ann. Math. Stat.* **38**, 759–770.

Neuts, M. F. (1979). A versatile Markovian Point process. *J. Appl. Prob.* **16**, 764–779.

Neuts, M. F. (1996). The single server queue with Poisson input and semi-Markov service times. *J. Appl. Prob.* **33**, 202–230.

Palm, C. (1943). Intensitat sschwankungen im Fernsprechverkehr. (Intensity fluctuations in telephone traffic.) *Erricsson Technics* **44**, 1–189.

Prabhu, N. U. (1960). Some results for the queue with Poisson arrivals. *J. Roy Stat. Soc.* **B22**, 104–107.

Prabhu, N. U. (1965). *Queues and Inventories*, Wiley, New York.

Prabhu, N. U., and Bhat, U. N. (1963a). Some first passage problems and their applications to queues. *Sankhyā* **A25**, 281–292.

Prabhu, N. U., and Bhat, U. N. (1963b). Further results for the queue with Poisson arrivals. *Opns. Res.* **11**, 380–386.

Ramaswami, V. (1980). The $N/G/1$ queue and its detailed analysis. *Adv. Appl. Prob.* **12**, 222–261.

Rego, V. (1988). Characterizations of equilibriun queue length distributions in $M/G/1$ queues. *Comp. & Opns. Res.* **15**, 7–17.

Rego, V., and Szpankowski, W. (1989). The presence of exponentiality in entropy maximized queues. *Comp. & Opns. Res.* **16**, 441–449.

Righter, R. (1999). A note on losses in $M/G/1/n$ queue. *J. Appl. Prob.* **36**, 1240–1243.

Rosenkrantz, W. A. (1983). Calculation of the L.T. of the length of the busy period for $M/G/1$ queue via martingales. *Ann. Prob.* **11**, 817–818.

Ross, S. M. (1980). *Introduction to Probability Models*, 2nd ed., Academic Press, New York.

Scott, M., and Ulmer, M. B., Jr. (1972). Some results for a simple queue with limited waiting room. *Zeit. F. Opns. Res.* **16**, 199–204.

Shanthikumar, J. G. (1988). DFR property of first-passage times and its preservation under geometric compounding. *Annals Prob.* **16**, 397–407.

Sobel, M. (1980). Some inequalities for multiserver queues. *Mgmt. Sci.* **26**, 951–956.

Sphicas, G. P., and Shimshak, D. G. (1978). Waiting time variability in some single server queueing systems. *J. Opnl. Res. Soc.* **29**, 65–70.

Stadje, W. (1985). The busy period of the queueing system $M/G/\infty$. *J. Appl. Prob.* **22**, 697–704.

Takács, L. (1962). *An Introduction to the Theory of Queues*, Oxford University Press, Oxford, UK.

Takács, L. (1963). A single server queue with feed-back. *Bell Sys. Tech. J.* **42**, 505–519.

Takács, L. (1967). *Combinatorial Methods in the Theory of Queues*, Wiley, New York.

Takács, L. (1969). On Erlang's formula. *Ann Math. Stat.* **40**, 71–78.

Takács, L. (1976). On the busy period of single-server queue with Poisson input and general service times. *Opns. Res.* **24**, 564–571.

Takagi, H. (1991). *Queueing Analysis, Vol. 1. Vacation and Priority Systems,* Part I, North Holland, Amsterdam.

Takagi, H. (1993). *Queueing Analysis, Vol. 2, Finite Systems,* North Holland, Amsterdam.

Takagi, H. (1996). A note on the response time in $M/G/1$ queue with service in random order and Bernoulli Feedback. *J. Opnl. Res. Soc. Japan* **39**, 486–500.

Takagi, H., and Lamarie, R. O. (1994). Busy period of an $M/G/1/K$ queue. *Opns. Res.* **42**, 192–193.

Takagi, H., and Sakamaki, Ken-ichi (1995). *Symbolic moment calculation for an $M/G/1$ queue.* Preprint #596, Univ. of Tsukuba, Japan.

Takine, T., Takagi, H., and Hasegawa, T. (1993). Analysis of an $M/G/1/K/N$ queue. *J. Appl. Prob.* **30**, 446–454.

Teghem, J., Loris-Teghem, J., and Lambotte, J. P. (1969). *Modéles d'attente $M/G/1$ et $GI/M/1$ à Arrivées et Services en Groupes.* Lecture Notes on O. R. 8, Springer-Verlag, Berlin.

Tijms, H. C. (1994). *Stochastic Models: An Algorithmic Approach,* Wiley, New York.

Tijms, H. C., and Hogenkamp, J. W. (1995). A Heuristic for the $M^X/G/c/c$ Loss Model in *Probability Models and Statistics: A. J. Medhi Festschrift,* pp. 69–72, New Age Int. (P) Ltd., Publ., New Delhi.

Trivedi, K. S. (1982). *Probability and Statistics with Reliability, Queueing and Computer Science Applications,* (2nd ed. 2001), Prentice-Hall.

Whitt, W. (1983). Comparing batch delays and customer delays. *Bell. Sys. Tech. J.* **62 (7)** 2001–2009.

Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Mgmt. Sci.* **38**, 708–723.

Willmot, G. E. (1988). A note on the equilibrum $M/G/1$ queuelength. *J. Appl. Prob.* **25**, 228–231.

Wolff, R. W., and Wrightson, C. W. (1976). An extension of Erlang's formula. *J. Appl. Prob.* **21**, 628–632.

Yadin, M., and Naor, P. (1963). Queueing systems with a removable service station. *Opnl. Res. Qrly.* **14**, 393–405.

Zhao, Y. (1994). Analysis of the $GI^X/M/c$ model. *Queueing Syst.* **15**, 347–364.

Queues with General Arrival Time and Service-Time Distributions

7.1 The $G/G/1$ Queue with General Arrival Time and Service-Time Distributions

In this system, one server provides service to customers one by one according to FIFO service discipline. Assume that the system is in steady state.

Let

t_n = instant of arrival of the nth customer

$u_n \equiv$ interarrival time between the nth and $(n+1)$th customer

$= t_{n+1} - t_n$

$v_n \equiv$ service time of the nth customer

$X_n \equiv v_n - u_n$

$W_n =$ waiting time in queue of the nth customer

$D_n \equiv$ instant of departure of the nth customer (and instant of commencement of service of the $(n+1)$th customer, if already in queue)

$\tau_n \equiv$ time between the nth and $(n+1)$th departures $= D_{n+1} - D_n$

$I_{n-1} \equiv$ idle time (if any) preceding the nth arrival (idle period, if any, preceding the nth arrival)

$J_n \equiv$ total idle period up to the instant of the nth arrival

$\equiv I_1 + I_2 + \cdots + I_n$

$I \equiv$ Idle period with DF $H(x) = Pr\{I \leq x\}$

$B =$ Busy period with DF $\beta(x) = Pr\{B \leq x\}$

$N \equiv$ Number served during a busy period

Assume further

(i) that $u_1, u_2, \ldots$ are IID random variables with a common DF $A(u) = P\{u_i \leq u\}$ with mean $E(u) = 1/\lambda$ and $\text{var}(u) = \sigma_u^2$;

(ii) that $v_1, v_2, \ldots$ are IID RV with a common DF $B(v) = P\{v_i \leq v\}$ with mean $E(v) = 1/\mu$ and $\text{var}(v) = \sigma_v^2$; and

(iii) that u_n and v_n are mutually independent. The RVs $X_n = v_n - u_n$ are IID. Let $K(x) = P\{X_n \leq x\}$ be its DF. Then

$$\alpha = E(X_n) = \frac{1}{\mu} - \frac{1}{\lambda} \quad \text{and} \quad \text{var}(X_n) = \sigma_u^2 + \sigma_v^2.$$

We shall assume that the system is in steady state. This happens *iff* $\rho = \lambda/\mu < 1$, which implies that

$$\alpha = \frac{1}{\mu} - \frac{1}{\lambda} < 0.$$

Let $S_n = X_1 + \cdots + X_n$. The random walk $\{S_n, n \geq 1\}$ is the basic process underlying the queueing model. We shall here be mainly interested in the stochastic process $\{W_n, n \geq 1\}$, a process in discrete time with continuous state space. Two cases arise.

First, a customer arrives to find the server busy. Suppose that the $(n+1)$th customer is such a customer. Its waiting time $W_{n+1}(\geq 0)$ is given by

$$\begin{aligned} W_{n+1} &= D_n - t_{n+1} \\ &= v_n + (D_{n-1} - t_{n+1}) \\ &= v_n + W_n - u_n \\ &= W_n + (v_n - u_n) = W_n + X_n(\geq 0). \end{aligned}$$

Second, a customer arrives to find the server idle. Suppose that the $(n+3)$th customer is such a customer. Its waiting time $W_{n+3} = 0$ since $D_{n+2} - t_{n+3} < 0$ or $W_{n+2} + v_{n+2} - u_{n+2} < 0$.

Thus, we can write

$$W_{n+1} = \begin{cases} W_n + v_n - u_n & \text{if } W_n + v_n - u_n \geq 0 \\ 0 & \text{if } W_n + v_n - u_n < 0. \end{cases} \tag{7.1.1}$$

Another expression of W_{n+1} is as follows:

$$W_{n+1} = \max(0, W_n + v_n - u_n). \tag{7.1.2}$$

The RV's v_n and u_n of the sequences $\{v_n\}$ and $\{u_n\}$ are independent among themselves and each other. The value of W_{n+1} depends on the sequence of RV's $W_i, i = 1, 2, \ldots, n$, only through its most recent value W_n plus a RV $X_n = v_n - u_n$, which is independent of all W_i for $i \leq n$. Thus, $\{W_n, n \geq 1\}$ is a Markov process with stationary transition probabilities.

Now $I_n = -\min(0, W_n + v_n - u_n)$; so that $I_n = 0$ or $I_n > 0$, and

$$I_n > 0 \Rightarrow I_n = I,$$

the idle period.

Note that in a $G/G/1$ system, the probability p_0 that the system is empty is $p_0 = 1 - \lambda E(v) = 1 - \rho$.

7.1.1 Lindley's integral equation

Let us denote the stationary distribution of W_n by

$$\lim_{n\to\infty} \{W_n \leq x\} = W(x). \tag{7.1.3}$$

We have

$$K(x) = P\{X_n \leq x\} = P\{v_n - u_n \leq x\}.$$

Conditioning on u_n we get

$$\begin{aligned} K(x) &= \int_{u=0}^{\infty} P\{v_n \leq x + u \mid u_n = u\} dA(u) \\ &= \int_{u=0}^{\infty} B(x + u) dA(u). \end{aligned} \tag{7.1.4}$$

We have, for $x \geq 0$,

$$\begin{aligned} W_{n+1}(x) &= P\{W_{n+1} \leq x\} \\ &= P\{W_n + X_n \leq x\}. \end{aligned}$$

Conditioning on W_n, we get

$$W_{n+1}(x) = \int_{0}^{\infty} P\{X_n \leq x - t \mid W_n = t\} dW_n(t). \tag{7.1.5}$$

Now since X_n is independent of W_n, we have

$$P\{X_n \leq x - t \mid W_n = t\} = P\{X_n \leq x - t\} = K(x - t).$$

Thus, for $x \geq 0$,

$$W_{n+1}(x) = \int_{0^-}^{\infty} K(x - t) dW_n(t).$$

Taking limits as $n \to \infty$ and noting that $\lim_{n\to\infty}\{W_n \le x\} = W(x)$, we get

$$W(x) = \int_{0^-}^{\infty} K(x-t)dW(t), \quad x \ge 0.$$

Further, we have

$$W(x) = 0, \quad x < 0.$$

Thus, we get Lindley's integral equation

$$\begin{aligned} W(x) &= \int_{0^-}^{\infty} K(x-t)dW(t), \quad 0 \le x < \infty \\ &= 0, \quad x < 0. \end{aligned} \tag{7.1.6}$$

The equation can be written in two other alternative forms as follows.

We have for $x \ge 0$,

$$\begin{aligned} W(x) &= K(x-t)\,W(t)|_{t=0}^{\infty} - \int_{0^-}^{\infty} dK(x-t)\,W(t) \\ &= \lim_{t\to\infty} K(x-t)\,W(t) - K(x-t)\,W(0^-) \\ &\quad - \int_{0^-}^{\infty} W(t)dK(x-t) \\ &= -\int_{0^-}^{\infty} W(t)dK(x-t). \end{aligned}$$

Thus,

$$\begin{aligned} W(x) &= -\int_{0^-}^{\infty} W(t)dK(x-t), \quad x \ge 0 \\ &= 0, \quad x < 0. \end{aligned} \tag{7.1.7}$$

Changing the variable in the integral by putting $x - t = u$, we get

$$\begin{aligned} W(x) &= \int_{-\infty}^{x} W(x-u)dK(u), \quad x \ge 0 \\ &= 0, \quad x < 0. \end{aligned} \tag{7.1.8}$$

Equations (7.1.6)–(7.1.8), which are all called Lindley's integral equation, are Wiener-Hopf integral equations. From the integral equation it is clear that the waiting-time distribution function $W(x)$ depends only on the distribution function $K(x) = Pr\{v_n - u_n \le x\}$—that is, on the DF of the difference of the service-time and interarrival-time distributions, rather than on the DFs of the individual distributions. This basic equation describes the waiting-time distribution of the $G/G/1$ queue. The Lindley integral equation, which looks

like a convolution integral, is not exactly of convolution type. The Lindley integral equation holds only for nonnegative values of the variables, while the distribution function vanishes for negative values of the variable. The integral equation can be solved using techniques of complex variable theory.

7.1.2 Laplace transform of *W*

Next we shall obtain the Laplace Transform of the waiting-time distribution as given in (7.1.8).

As the integral on the RHS does not have a numerical value (other than zero) for negative x, $(x < 0)$, we shall define an integral as follows for negative values of x.

$$W^-(x) = \int_{-\infty}^{x} W(x-u)dK(u) \quad \text{when } x < 0$$
$$= 0 \quad \text{when } x \geq 0. \tag{7.1.9}$$

Combining (7.1.8) and (7.1.9), we get

$$W^-(x) + W(x) = \int_{-\infty}^{x} W(x-u)dK(u), \quad -\infty < x < \infty, \tag{7.1.10}$$

which holds for all real x.

We define the two-sided Laplace Transforms of $W(t)$ and $W^-(t)$ as follows:

$$\tilde{W}(s) = \int_{-\infty}^{\infty} e^{-st}W(t)dt = \int_{0}^{\infty} e^{-st}W(t)dt$$

$$\tilde{W}^-(s) = \int_{-\infty}^{\infty} e^{-st}W^-(t)dt = \int_{-\infty}^{0} e^{-st}W^-(t)dt.$$

Let $A^*(s)$ and $B^*(s)$ be the LST of u_n and v_n, respectively. Then since K is the distribution function of $X_n = v_n - u_n$, the two-sided LST of $K(u)$ is given by

$$K^*(s) = \int_{-\infty}^{\infty} e^{-st}dK(t) = B^*(s)\,A^*(-s). \tag{7.1.11}$$

Taking the two-sided LT of the RHS of (7.1.10), we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{x} e^{-sx}W(x-u)dK(u)$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{x} e^{-(x-u)s}W(x-u)e^{-us}\,dK(u)$$
$$= \left[\int_{-\infty}^{\infty} e^{-(x-u)s}W(x-u)du\right]\left[\int_{-\infty}^{\infty} e^{-us}\,dK(u)\right]$$
$$\text{(since } W(x-u) = 0 \quad \text{for } u \geq x\text{)}$$

$$= \left[\int_{-\infty}^{\infty} e^{-ts} W(t)\,dt\right]\left[\int_{-\infty}^{\infty} e^{-us}\,dK(u)\right]$$
$$= \bar{W}(s)K^*(s). \tag{7.1.12}$$

Using (7.1.11) and (7.1.12)

$$\bar{W}(s) + \bar{W}^-(s) = \bar{W}(s)K^*(s)$$
$$= \bar{W}(s)B^*(s)A^*(-s),$$

whence, we get

$$\bar{W}(s) = \frac{\bar{W}^-(s)}{B^*(s)A^*(-s) - 1} \tag{7.1.13}$$

Thus, the LT $\bar{W}(s)$ of the waiting time W is obtained in terms of the LST of the RVs u_n and v_n and the unknown $\bar{W}^-(s)$. Solution of Lindley's equation becomes difficult because determination of the unknown $\bar{W}^-(s)$ requires techniques of complex variable theory.

Example 7.1. The $M/M/1$ queue

Here

$$A^*(s) = \frac{\lambda}{\lambda + s}, \qquad B^*(s) = \frac{\mu}{\mu + s},$$

so that the Eq. (7.1.13) becomes

$$\bar{W}(s) = \frac{\bar{W}^-(s)}{\frac{\lambda}{\lambda - s}\frac{\mu}{\mu + s} - 1}$$
$$= \frac{(\lambda - s)(\mu + s)\bar{W}^-(s)}{s(\mu - \lambda + s)}. \tag{7.1.14}$$

Now to find $\bar{W}^-(s)$, we note that

$$W^*(\lambda) = \int_0^{\infty} e^{-\lambda t}\,dW(t) = e^{-\lambda t}W(t)|_{t=0}^{\infty} + \lambda\int_0^{\infty} e^{-\lambda t}W(t)dt.$$

We are concerned with computation of $W(t)$ for $t > 0$; we may assume for the time being that $W(0) = 0$. We can ultimately put $W(0) = p_0 = 1 - \rho$, which holds for all $M/G/1$ queues. Thus,

$$W^*(\lambda) = \lambda\bar{W}(\lambda).$$

Now $\int_0^{\infty} e^{-\lambda t}\,dW(t)$ is the probability that no customer arrives during the waiting time of an arbitrary customer. This is the probability that an arrival finds no other customer waiting for service or that the arrival finds the system idle or with one customer in service.

Thus, for an $M/M/1$ system,

$$W^*(\lambda) = \int_0^\infty e^{-\lambda t}\, dW(t) = p_0 + p_1 = (1-\rho) + \rho(1-\rho)$$
$$= (1-\rho)(1+\rho),$$

so that

$$\bar{W}(\lambda) = \frac{(1-\rho)(1+\rho)}{\lambda} \neq 0. \tag{7.1.15}$$

Since $\bar{W}(\lambda) \neq 0$, it is clear from (7.1.14) that $\bar{W}^-(s)$ is of the form

$$\bar{W}^-(s) = \frac{c}{\lambda - s} \quad \text{where } c \text{ is a constant or}$$
$$(\lambda - s)\bar{W}^-(s) = c.$$

From (7.1.14) we find

$$\bar{W}(\lambda) = \lim_{s\to\lambda}\left[\frac{c(\mu+s)}{s(\mu-\lambda+s)}\right]$$
$$= c\frac{\mu+\lambda}{\lambda\mu} = \frac{c}{\lambda}(1+\rho).$$

Using (7.1.15), we get

$$\frac{(1-\rho)(1+\rho)}{\lambda} = \frac{c}{\lambda}(1+\rho) \quad \text{or}$$
$$c = 1-\rho.$$

Thus,

$$\bar{W}^-(s) = \frac{1-\rho}{\lambda - s}. \tag{7.1.16}$$

Substituting this value of $\bar{W}^-(s)$ in (7.1.14), we get

$$\bar{W}(s) = \frac{(\lambda - s)(\mu + s)}{s(\mu - \lambda + s)} \cdot \frac{1-\rho}{\lambda - s}$$
$$= \frac{(\mu+s)(1-\rho)}{s(\mu-\lambda+s)}$$
$$= \frac{1-\rho}{s} + \frac{\lambda(1-\rho)}{s(\mu-\lambda+s)}. \tag{7.1.17}$$

Inverting the LT, we get

$$W(x) = (1-\rho) + \frac{\lambda(1-\rho)}{\mu-\lambda}\left[1 - e^{-(\mu-\lambda)x}\right]$$
$$= 1 - \rho e^{-\mu(1-\rho)x}, \quad (x > 0). \tag{7.1.18}$$

We now note that

$$W(0) = p_0 = 1 - \rho$$

Note that

$$\tilde{W}(s) = \frac{(1-\rho)(\mu+s)}{s(\mu-\lambda+s)}$$

is the Pollaczek-Khinchin formula for $M/G/1$ for the special case of exponential service time, $G \equiv M$. Result (7.1.18) was earlier obtained directly.

7.1.3 Generalization of the Pollaczek-Khinchin transform formula

We shall now obtain the LST of W in terms of the LST of the idle time I and the LST of X_n. We have

$$W_{n+1} - I_n = W_n + X_n,$$

with $W_{n+1} I_n = 0$, and $I_n = I$ whenever $I_n > 0$. Now since W_n and X_n are independent, we have

$$E\left[e^{-s(W_{n+1}-I_n)}\right] = E\left[e^{-sW_n}\right] E\left[e^{-sX_n}\right]. \tag{7.1.19}$$

Again, since $W_{n+1} = 0$, when $I_n > 0$ and $W_n \neq 0$, when $I_n = 0$

$$\begin{aligned} E\left[e^{-s(W_{n+1}-I_n)}\right] &= E\left[e^{-s(-I_n)} \mid I_n > 0\right] Pr(I_n > 0) \\ &\quad + E\left[e^{-sW_{n+1}} \mid I_n = 0\right] Pr(I_n = 0). \end{aligned} \tag{7.1.20}$$

Now

$$\begin{aligned} E\left[e^{-sW_{n+1}}\right] &= E\left[e^{-sW_{n+1}} \mid I_n = 0\right] Pr(I_n > 0) \\ &\quad + E\left[e^{-sW_{n+1}} \mid I_n > 0\right] Pr(I_n > 0). \end{aligned} \tag{7.1.21}$$

Since $W_{n+1} = 0$ when $I_n > 0$,

$$E\left[e^{-sW_{n+1}} \mid I_n > 0\right] = 1,$$

and since $Pr(I_n > 0) = Pr\{\text{arrivals find system idle}\} = a_0$

$$E\left[e^{-sW_{n+1}}\right] = E\left[e^{-sW_{n+1}} \mid I_n = 0\right] Pr(I_n = 0) + a_0$$

and from (7.1.20), we get

$$\begin{aligned} E\left[e^{-s(W_{n+1}-I_n)}\right] &= E\left[e^{-s(I_n)} \mid I_n > 0\right] Pr(I_n > 0) \\ &\quad + E\left[e^{-sW_{n+1}}\right] - a_0. \end{aligned} \tag{7.1.22}$$

Thus, from (7.1.19) and (7.1.22), we get

$$E\left[e^{-sW_n}\right]E\left[e^{-sX_n}\right] = E\left[e^{-s(-I_n)} \mid I_n > 0\right]a_0 + E\left[e^{-sW_{n+1}}\right] - a_0. \quad (7.1.23)$$

Note that when $I_n > 0$, $I_n = I$ and in steady state

$$\lim E\left[e^{-sW_n}\right] = \lim E\left[e^{-sW_{n+1}}\right] = E\left[e^{-sW}\right].$$

Thus, taking limit of (7.1.23), we get

$$E\left[e^{-sW}\right]E\left[e^{-sX_n}\right] = E\left[e^{-s(-I)}\right]a_0 + E\left[e^{-sW}\right] - a_0. \quad (7.1.24)$$

Now

$$E\left[e^{-sW}\right] = W^*(s)$$

is the LST of W. $E[e^{-sX_n}] = K^*(s)$ is the two-sided Laplace-Stieltjes Transform of

$$K(u) = P(X_n = v_n - u_n \leq u);$$
$$E\left[e^{-s(I)}\right] = I^*(s)$$

is the Laplace-Stieltjes Transform of I. Thus, from (7.1.24) we get

$$W^*(s) = \frac{a_0[1 - I^*(-s)]}{1 - K^*(s)}. \quad (7.1.25)$$

The preceding generalization of the Pollaczek-Khinchin (Marshall, 1968c) formula holds for the $G/G/1$ system.

Example 7.2. The $M/G/1$ system

For a Poisson arrival system, the idle-time distribution is the same as the interarrival-time distribution, so that $A^*(s) = I^*(s) = \lambda/(\lambda + s)$. Again, $a_0 = p_0 = 1 - \rho$. Further,

$$\begin{aligned} K^*(s) &= E\left[e^{-s(v_n - u_n)}\right] \\ &= E[e^{-sv_n}]E[e^{su_n}] \\ &= B^*(s)\frac{\lambda}{\lambda - s}, \end{aligned}$$

where $B^*(s)$ is the LST of the service-time distribution. Substituting in (7.1.25) we get

$$\begin{aligned} W^*(s) &= \frac{(1-\rho)\left[\frac{-s}{(\lambda - s)}\right]}{1 - \left[\frac{\lambda}{(\lambda - s)}\right]B^*(s)} \\ &= \frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)}, \end{aligned} \quad (7.1.26)$$

which is exactly the Pollaczek-Khinchin Transform formula for the LST of waiting time.

Noting that $s\bar{W}(s) = W^*(s)$, we get the P-K formula for the LST of the waiting time.

Example 7.3. The $M/M/1$ system
For an $M/M/1$ system also

$$p_0 = 1 - p \quad \text{and}$$
$$A^*(s) = I^*(s) = \frac{\lambda}{(\lambda + s)}.$$

Since $B^*(s) = \mu/(\mu + s)$, we get from (7.1.26)

$$W^*(s) = \frac{s(1-\rho)}{s - \lambda + \frac{\lambda\mu}{(\mu+s)}}$$
$$= \frac{(1-\rho)(\mu+s)}{\mu + s - \lambda}. \tag{7.1.27}$$

Noting that $s\bar{W}(s) = W^*(s)$, we get the same result as obtained in (7.1.17).

7.2 Mean and Variance of Waiting Time *W*

7.2.1 Mean of *W* (single-server queue)

Marshall (1968a) obtained the mean and the variance of W. The mean of the waiting time can be obtained in terms of the mean and variance of the inter-arrival, service, and idle-time distributions. This is given shortly in Theorem 7.1.

First, we prove the following.
For a $G/G/1$ queue with $\rho < 1$, the mean idle time $E(I)$ is given by

$$E(I) = \frac{\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)}{a_0} = \frac{(1-\rho)}{\lambda a_0}, \tag{7.2.1}$$

where a_0 is the probability that an arrival finds the system empty.

We have

$$W_{n+1} - I_n = W_n + X_n. \tag{7.2.2}$$

Since the queue is stationary, $E(W_{n+1}) = E(W_n)$. Whenever $I_n > 0$, I_n equals the idle period I. We have

$$E(I_n) = Pr\{\text{system is found empty}\} \times E\{\text{idle time}\}$$
$$= a_0 E(I).$$

Thus,

$$-a_0 E(I) = E(X_n) = E(v_n) - E(u_n),$$

whence

$$E(I) = \frac{\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)}{a_0} = \frac{(1-\rho)}{\lambda a_0}.$$

Notes:

(1) Either W_{n+1} or I_n is always zero, so that

$$W_{n+1} I_n = 0, \qquad W_{n+1}^2 I_n = 0, \quad \text{and}$$
$$W_{n+1} I_n^2 = 0.$$

(2) Let B be the busy period. Then $\{B\}$ and $\{I\}$ are alternating renewal processes and

$$\frac{E(I)}{E(B) + E(I)} = 1 - \rho.$$

Using (7.2.1), we get

$$E(B) = \frac{1}{\mu a_0}.$$

(3) For the Poisson input system, $E(I) = I/\lambda, a_0 = p_0 = 1 - \rho$, and

$$E(B) = \frac{1}{\mu(1-\rho)}.$$

Theorem 7.1. *For all G/G/1 queues with $\rho < 1$,*

$$E(W) = \frac{\lambda^2(\sigma_u^2 + \sigma_v^2) + (1-\rho)^2}{2\lambda(1-\rho)} - \frac{v_h^{(2)}}{2v_h}, \tag{7.2.3}$$

where v_h and $v_h^{(2)}$ are the first and second moments of the idle period I.

Proof: Squaring (7.2.2) and noting that

$$W_{n+1} \cdot I_n = 0,$$

we get

$$W_{n+1}^2 + I_n^2 = W_n^2 + 2W_n X_n + X_n^2. \tag{7.2.4}$$

We have

$$\begin{aligned} E(I_n^2) &= Pr\{\text{system is empty}\}\, E(I^2) \\ &= a_0 E(I^2) \quad \text{and} \\ E(W_n X_n) &= E(W_n) E(X_n), \end{aligned}$$

since W_n, X_n are independent. Thus, taking the expectation of (7.2.4) and noting that in steady state $E(W_n) = E(W)$ and $E(W_{n+1}^2) = E(W_n^2)$, we get

$$a_0 E(I^2) = E(X_n^2) + 2E(W)E(X_n),$$

so that

$$E(W) = \frac{a_0 E(I^2) - E(X_n^2)}{2E(X_n)}. \tag{7.2.5}$$

Now

$$\begin{aligned} E(X_n) &= \frac{1}{\mu} - \frac{1}{\lambda} = \frac{1}{\lambda}(\rho - 1), \\ E(X_n^2) &= E(v_n^2 + u_n^2 - 2u_n v_n) = E(v_n^2) + E(u_n^2) - 2E(u_n)E(v_n) \\ &\qquad \text{(since } u_n, v_n \text{ are independent)} \\ &= \sigma_v^2 + \frac{1}{\mu^2} + \sigma_u^2 + \frac{1}{\lambda^2} - 2\left(\frac{1}{\lambda}\right)\left(\frac{1}{\mu}\right) = \sigma_u^2 + \sigma_v^2 + \frac{(1-\rho)^2}{\lambda^2}, \\ v_h &= E(I) = \frac{\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)}{a_0} = \frac{1}{\lambda a_0}(1-\rho) = -\frac{E(X_n)}{a_0} \quad \text{and} \\ v_h^{(2)} &= E(I^2). \end{aligned}$$

From (7.2.5) we get

$$E(W) = \frac{E(X_n^2)}{-2E(X_n)} - \frac{E(I^2)}{2E(I)}, \tag{7.2.5a}$$

and using the preceding relations, we get

$$\begin{aligned} E(W) &= \frac{a_0 v_h^{(2)} - \left[\sigma_u^2 + \sigma_v^2 + \frac{(1-\rho)^2}{\lambda^2}\right]}{(2/\lambda)(\rho - 1)} \\ &= \frac{\lambda^2(\sigma_u^2 + \sigma_v^2) + (1-\rho)^2}{2\lambda(1-\rho)} - \frac{v_h^{(2)}}{2v_h}, \end{aligned} \tag{7.2.5b}$$

which is the relation (7.2.3). ■

Notes:

(1) Denoting

$$C_g \equiv \text{ coeff. of variation of } X_n$$

we can put (7.2.3) in an alternative form in terms of C_g.

(2) The quantity $v_h^{(2)}/2v_h$ is the expectation of the RV Z with distribution function

$$\frac{1}{v_h}\int_0^t [1 - H(u)]du,$$

where $H(u)$ is the DF of the idle period I. Z is the equilibrium residual idle period.

(3) For a queue with Poisson arrivals, we have

$$v_h = E(I) = \frac{1}{\lambda} \quad \text{and} \quad v_h^{(2)} = E(I^2) = \frac{2}{\lambda^2}, \quad \sigma_u^2 = \frac{1}{\lambda^2},$$

so that for an $M/G/1$ queue, Eq. (7.2.3) becomes

$$\begin{aligned} E(W) &= \frac{1 + \lambda^2\sigma_v^2 + 1 - 2\rho + \rho^2}{2\lambda(1-\rho)} - \left(\frac{2}{\lambda^2}\right)\Big/\left(\frac{2}{\lambda}\right) \\ &= \frac{\rho^2 + \lambda^2\sigma_v^2}{2\lambda(1-\rho)}, \end{aligned} \tag{7.2.6}$$

which is the mean waiting time in an $M/G/1$ queue (Eq. 6.3.12).

(4) For the $D/D/1$ queue,

$$\sigma_u^2 = \sigma_v^2 = 0 \quad \text{and}$$
$$v_h^{(2)} = (v_h)^2 = \left(\frac{1}{\lambda} - \frac{1}{\mu}\right)^2,$$

so that (7.2.3) reduces to

$$E(W) = 0.$$

7.2.2 Variance of W

Theorem 7.2. *For a $G/G/1$ queue with $\rho < 1$ with FIFO service discipline, the variance σ_W^2 of W is given by*

$$\sigma_W^2 = \frac{3E(X_n^3)}{-3E(X_n)} + \left[\frac{E(X_n^2)}{-2E(X_n)}\right]^2 + \frac{E(I^3)}{3E(I)} - \left[\frac{E(I^2)}{2E(I)}\right]^2 \tag{7.2.7}$$

$$= \frac{\left[\lambda\left(v_n^{(3)} - v_v^{(3)}\right) - 3\left(\rho v_u^{(2)} - v_v^{(2)}\right)\right]}{3(1-\rho)} + \left[\frac{\left\{\lambda^2\left(\sigma_u^2 + \sigma_v^2\right) + (1-\rho)\right\}^2}{2\lambda(1-\rho)}\right]^2 + \frac{v_h^{(3)}}{3v_h} - \left[\frac{v_h^{(2)}}{2v_h}\right]^2, \quad (7.2.8)$$

where $v_u^{(r)}, v_u^{(r)}$, *and* $v_h^{(r)}$ *are the rth moments* $(r = 1, 2, 3)$ *of the interarrival, service-time, and idle-time distributions, respectively, and* $v^{(1)} = v$.

Proof: Cubing both sides of (7.2.2)

$$W_{n+1} - I_n = W_n + X_n,$$

noting that $I_n^2 W_{n+1} = I_n W_{n+1}^2 = 0$ and taking expectations, we get

$$E\left[W_{n+1}^3 - I_n^3\right] = E\left[W_n^3 + X_n^3 + 3W_n^2 X_n + 3W_n X_n^2\right]. \quad (7.2.9)$$

In steady state

$$E\left[W_{n+1}^r\right] = E\left[W_n^r\right] = E[W^r], \quad r = 1, 2, 3.$$

Noting that W_n and X_n are independent, we get from (7.2.9)

$$E\left[-I_n^3\right] = E\left[X_n^3\right] + 3E[X_n]E[W^2] + 3E[W]E\left[X_n^2\right]. \quad (7.2.10)$$

Using $E(I_n^3) = a_0 E(I^3)$ and $E(X_n) = -a_0 E(I)$, we get

$$E(W^2) = \frac{E\left(X_n^3\right)}{-3E(X_n)} + \frac{E\left(X_n^2\right)}{-E(X_n)} E(W) + \frac{E(I^3)}{2E(I)}. \quad (7.2.11)$$

Putting the value of

$$E(W) = \frac{E\left(X_n^2\right)}{-2E(X_n)} - \frac{E(I^2)}{2E(I)}$$

from (7.2.5a), we get

$$\begin{aligned}\sigma_W^2 &= E(W^2) - [E(W)]^2 \\ &= \frac{E\left(X_n^3\right)}{-3E(X_n)} + \frac{E(I^3)}{3E(I)} + \left[\frac{E\left(X_n^2\right)}{-E(X_n)} + \frac{E\left(X_n^2\right)}{2E(X_n)} + \frac{E(I^2)}{2E(I)}\right] \\ &\quad \times \left[\frac{E\left(X_n^2\right)}{-2E(X_n)} - \frac{E(I^2)}{2E(I)}\right] \\ &= \frac{E\left(X_n^3\right)}{-3E(X_n)} + \left[\frac{E\left(X_n^2\right)}{-2E(X_n)}\right]^2 + \frac{E(I^3)}{3E(I)} - \left[\frac{E(I^2)}{2E(I)}\right]^2,\end{aligned}$$

which gives (7.2.7). Now

$$E(X_n) = -\frac{1}{\lambda}(1-\rho)$$

$$E(X_n^2) = \frac{\left[\lambda^2(\sigma_u^2+\sigma_v^2)+(1-\rho)^2\right]}{\lambda^2}$$

$$E(X_n^3) = v_u^{(3)} - v_v^{(3)} + 3v_u^{(2)}\left(\frac{1}{\mu}\right) + 3\left(\frac{1}{\lambda}\right)v_v^{(2)}$$

$$= \frac{\left[\lambda\left(v_u^{(3)}-v_v^{(3)}\right) - 3\left(\rho v_u^{(2)} - v_v^{(2)}\right)\right]}{\lambda}.$$

Using these expressions in (7.2.7), we at once get (7.2.8). ■

Note: It is tacitly assumed in the preceding derivations that all moments up to the required order (up to order two for validity of (7.2.4) and up to order three for validity of (7.2.8)) of the three distributions exist. Marshall (1968c) shows that the necessary and sufficient condition for the existence of all the moments up to order three is that the first three moments of the interarrival- and service-time distributions exist. For approximations of waiting time in a $GI/G/1$ queue, see Fredericks (1982) and Jagerman (1987).

7.2.3 Multiserver queues: approximation of mean waiting time

We discuss here one of the main performance measures of multiserver queueing systems in steady state: the mean waiting (queueing) time. While exact and closed-form expressions are known for simpler systems $M/M/c$, $M/G/1$, no such result is available in case of more general systems such as $M/G/c$, $G/G/c$. Various approximations have been suggested for obtaining mean waiting times for such systems. The approximations are in closed forms that combine analytical solutions of simpler systems, and so approximations are known as system interpolations.

One thing that facilitates use of such approximations is that tables for building block systems ($M/M/c$, $M/D/c$, $D/M/c$) for specific values of the parameters are available (e.g., Page, 1972, 1982; Seelen *et al.* 1985). We indicate below some of these approximations. Denote the mean waiting (queueing) time of a $A/B/c$ system in steady state by $EW(A/B/c)$. Let c_a^2, c_v^2 be the squares of coefficient of variation for the interarrival and service-time distributions, respectively. We recall that, by virtue of the Pollaczek-Khinchin formula

$$EW(M/G/1) = \frac{1+c_v^2}{2}\,EW(M/M/1). \tag{7.2.12}$$

Based on this, the following heuristic approximation for the c-server system has been suggested:

$$E\,W(M/G/c) \simeq \frac{1+c_v^2}{2}\,E\,W(M/M/c). \tag{7.2.13}$$

(Lee and Longton, 1957)

The mean waiting time for both $M/M/c$ and $M/G/1$ systems can be written as

$$\begin{aligned} E\,W(M/M/c) &= E\,W(M/G/1) \\ &= \frac{C(c,a)}{c\mu(1-\rho)} = \frac{1+c_v^2}{2}\,\frac{\rho C(c,a)}{\lambda(1-\rho)}. \end{aligned}$$

See (3.6.21) and (6.3.12c); note that $C(c,a) = \rho$ when $c = 1$ (see (3.6.4)).

Based on this, an approximation suggested (Hokstad, 1978; Stoyan, 1976; Nozaki and Ross, 1978; Maaloe, 1973), is

$$E(M/G/c) \simeq \frac{1+c_v^2}{2\lambda(1-\rho)}\,\rho C(c,a).$$

The approximation appears quite satisfactory for large ρ. Harel and Zipkin (1987) show that $f(\rho) = 1/\{E(W) + E(v)\}$ is strictly concave for $M/G/c$, $c \geq 2$ for sufficiently light traffic.

From (6.3.12), we have also the closed-form expression

$$E\,W(M/G/1) = c_v^2\,E\,W(M/M/1) + \left(1-c_v^2\right)E\,W(M/D/1). \tag{7.2.14}$$

Based on this, Björklund and Elldin (1964) suggested the following approximation for the c-server case:

$$E\,W(M/G/c) \simeq c_v^2\,E\,W(M/M/c) + \left(1-c_v^2\right)E\,W(M/D/c). \tag{7.2.15}$$

in terms of those of $M/M/c$ and $M/D/c$. Kimura (1986) suggested the approximation

$$E\,W(M/G/c) \simeq \frac{1+c_v^2}{\frac{2c_v^2}{E\,W(M/M/c)} + \frac{1-c_v^2}{E\,W(M/D/c)}}. \tag{7.2.16}$$

For $E\,W(M/D/c)$, approximations given by Cosmetatos (1975, 1976) (or refinements thereof: Kimura, 1991) could be used.

Approximations have also been provided by Boxma *et al.* (1979), Tijms (1987), Tijms (1994) and Tijms *et al.* (1981).

Approximations have also been suggested for queue-length and waiting-time distributions—for example, Kimura (1986, 1991, 1993), Tijms *et al.* (1981), and Van Hoorn and Tijms (1982).

For general interarrival and service times, one has the Kingman (1962) approximation for heavy traffic ($\rho \to 1 - 0$; see Section 8.1)

$$E\,W(G/G/1) \simeq \frac{c_a^2 + c_v^2}{2} E\,W(M/M/1). \tag{7.2.17}$$

Krämer and Langenbach-Belz (1976) suggested

$$E\,W(G/G/1) \simeq \frac{c_a^2 + c_v^2}{2} kE\,W(M/M/1), \tag{7.2.18}$$

where

$$k \equiv k(\rho, c_a^2, c_v^2) = \begin{cases} \exp\left\{-\frac{2(1-\rho)}{3\rho}\frac{(1-c_a^2)^2}{c_a^2+c_v^2}\right\}, & c_a^2 \le 1 \\ \exp\left\{-(1-\rho)\frac{c_a^2-1}{c_a^2+c_v^2}\right\}, & c_a^2 > 1. \end{cases} \tag{7.2.19}$$

This approximation (7.2.18) could be taken as a refinement of Kingman's heavy traffic approximation (7.2.17). As a natural extension of Kingman's result (7.2.17), an approximation suggested (Kimura) for $G/G/c$ queue is

$$E\,W(G/G/c) \simeq \frac{c_a^2 + c_v^2}{2} E\,W(M/M/c). \tag{7.2.20}$$

Page (1982) suggested the approximation

$$\begin{aligned} E\,W(G/G/c) \simeq c_a^2\, c_v^2 E\,W(M/M/c) + c_a^2\left(1 - c_v^2\right) E\,W(M/D/c) \\ + \left(1 - c_a^2\right) c_v^2 E\,W(D/M/c). \end{aligned} \tag{7.2.21}$$

This approximation coincides with (7.2.15) for the case $M/G/c$. Kimura (1986) also suggested the approximation,

$$E\,W(G/G/c) \simeq \frac{k\left(c_a^2 + c_v^2\right)}{\frac{2\left(c_a^2+c_v^2-1\right)}{E\,W(M/M/c)} + \frac{1-c_v^2}{E\,W(M/D/c)} + \frac{k_{01}\left(1-c_a^2\right)}{E\,W(D/M/c)}}, \tag{7.2.22}$$

$$c_a^2 \le 1$$

$$\simeq \left(c_a^2 + c_v^2 - 1\right) E\,W(M/M/c) + \left(1 - c_v^2\right) E\,W(M/D/c) + \frac{1 - c_a^2}{k_{01}} E\,W(D/M/c), \quad c_a^2 > 1, \tag{7.2.23}$$

where k is given by (7.2.19) and

$$k_{01} \equiv k(\rho, 0, 1) = \exp\{-2(1-\rho)/3\rho\}. \tag{7.2.24}$$

Approximations for $G/G/c$ have also been provided by Shore (1988a,b). For a discussion (with some comparisons with numerical results), refer to the survey paper by Kimura (1994).

7.3 Queues with Batch Arrivals $G^{(X)}/G/1$

Marshall (1968b) extended his results to queues with batch arrival and batch service.

Suppose that the arrivals in batches occur in a renewal process with rate λ and each arrival consists of a batch of customers of random size X with probability mass function

$$P(X = m) = a_m, \quad m = 0, 1, 2, \ldots$$

with rth moment

$$v_X^{(r)} \quad (v_X^{(1)} = v_X), \quad r = 1, 2, \ldots.$$

Customers arriving in a batch are assumed to be numbered in some way to denote the order of service, and service discipline is FIFO.

Let $V_{k,n}$ be the service time of the kth customer in the nth batch and let $V_{k,n}$ be IID RVs with common DF $A(\cdot)$ for all n and $k \geq 1$ and $V_{0,n} \equiv 0$. Let $W_{1,n}$ be the waiting time in queue of the first customer served of the nth batch. Let

$$V_n^* = V_{0,n} + V_{1,n} + \cdots + V_{X,n} \tag{7.3.1}$$

be the total service time of all arrivals of the nth batch and let its DF be $B^*(v) = P(V_n^* \leq v)$.

Now $\{W_{1,n+1}\}$ satisfies the relation

$$W_{1,n+1} = \max[0, \ W_{1,n} + X_n^*], \tag{7.3.2}$$

where $X_n^* = V_n^* - T_n^*$, T_n^* being the interarrival time between the arrival of the nth and $(n+1)$th batches. Assume that

$$\lambda v_X < \mu \left(\rho = \frac{\lambda v_X}{\mu} < 1 \right),$$

the queue is stationary, and

$$E[W_{1,n}] \to E[W_{1,\infty}],$$

the mean waiting time in queue of the first customer of an arbitrary batch.

We shall use the relation for the $G/G/1$ queue with single-arrival and individual service,

$$E[W] = \frac{E[X_n]^2}{-2E(X_n)} - \frac{E(I^2)}{2E(I)}$$

to find the mean waiting time in queue for the first customer in each batch.

For the batch-arrival case, we have V_n^* as the sum of a random number of IID RV each with mean v_v and variance σ_v^2. Thus,

$$\begin{aligned}\sigma_{V^*}^2 &= E(X)\,\mathrm{var}(v) + \mathrm{var}(X)[E(v)]^2 \\ &= v_X\sigma_v^2 + \sigma_X^2\left(\frac{1}{\mu^2}\right).\end{aligned} \tag{7.3.3}$$

Thus, applying the result of Theorem 7.1, we get

$$E[W_{1,\cdot}] = \frac{\sigma_u^2 + \sigma_{V^*}^2}{\frac{2}{\lambda}\left(1 - \frac{\lambda v_X}{\mu}\right)} + \frac{1 - \frac{\lambda v_X}{\mu}}{2\lambda} - \frac{v_h^{(2)}}{2v_h}. \tag{7.3.4}$$

We now proceed to find the mean waiting time in queue of an *arbitrary* customer in a batch. The mean total additional waiting time of all customers in an average batch is obtained, and it is divided by the average number of customers per batch. Let Z_n be the total additional waiting time of all customers in some batch with n arrivals. The $Z_n = 0$ if $X = 0, 1$ and for $X \geq 2$

$$Z_n = V_{1,n} + (V_{1,n} + V_{2,n}) + \cdots + \left(V_{1,n} + V_{2,n} + \cdots + V_{(X-1),n}\right),$$

so that

$$\begin{aligned}E(Z_n \mid X = k) &= E[(V_{1,n}) + (V_{1,n} + V_{2,n}) + \cdots + (V_{1,n} + \cdots + V_{k-1,n})] \\ &= \frac{k(k-1)}{2}E(V) = \frac{k(k-1)1}{2}\,\frac{1}{\mu} \quad \text{for } k \geq 0\end{aligned}$$

and thus,

$$\begin{aligned}E(Z_n) &= \sum_{k=1}^{\infty} E(z_n \mid X = k)Pr(X = k) \\ &= \sum_{k=1}^{\infty} \frac{k(k-1)}{2}\,\frac{1}{\mu}\,a_k \\ &= \frac{1}{2\mu}\left[\sum_{k=1}^{\infty} k^2 a_k - \sum_{k=1} k a_k\right] \\ &= \frac{1}{2\mu}\left[v_X^{(2)} - v_X\right] \quad \text{and} \\ \frac{E(Z_n)}{\text{average number of customers per batch}} &= \frac{E(Z_n)}{v_X} \\ &= \frac{1}{2\mu}\left[\frac{v_X^{(2)}}{v_X} - 1\right].\end{aligned} \tag{7.3.5}$$

Thus, the expected wait of any unspecified customer $E(W)$ is then obtained by using (7.3.4) and (7.3.5). We have

$$E(W) = E(W_{1,.}) + \frac{1}{2\mu}\left[\frac{v_X^{(2)}}{v_X} - 1\right]. \tag{7.3.6}$$

7.4 The Output Process of a $G/G/1$ System

The interdeparture interval τ_n between the nth and $(n+1)$ departures is given by

$$\begin{aligned}\tau_n &= D_{n+1} - D_n \\ &= t_{n+1} + W_{n+1} + v_{n+1} - (t_n + W_n + v_n) \\ &= (t_{n+1} - t_n) + W_{n+1} - W_n + v_{n+1} - v_n. \end{aligned} \tag{7.4.1}$$

In steady state, $E(W_{n+1}) = E(W_n)$ so that the expected interdeparture interval is given by

$$\begin{aligned}E(\tau_n) &= E(t_{n+1} - t_n) \\ &= E(u_n) = \frac{1}{\lambda}. \end{aligned} \tag{7.4.2}$$

Note: For a single-server system in equilibrium, the departure rate is equal to the arrival rate.

Theorem 7.3. *The variance of the interdeparture interval $\tau_n \equiv \tau$ in a $G/G/1$ queue is given by*

$$\text{var}(\tau) = \sigma_v^2 - \frac{(1-\rho)^2}{\lambda^2} + \left[\frac{(1-\rho)}{\lambda}\right]\left[\frac{v_h^{(2)}}{v_h}\right].$$

Proof: We have

$$\tau_n = v_{n+1} + I_n, \tag{7.4.3}$$

where v_{n+1} and I_n are independent. Hence

$$\text{var}(\tau_n) = \text{var}(v_{n+1}) + \text{var}(I_n). \tag{7.4.4}$$

Again,

$$\begin{aligned} W_{n+1} - I_n &= W_n + X_n = W_n + u_n - v_n \quad \text{and} \\ \text{var}(W_{n+1} - I_n) &= \text{var}(W_n) + \text{var}(u_n) + \text{var}(v_n) \\ &= \sigma_W^2 + \sigma_u^2 + \sigma_v^2. \end{aligned} \tag{7.4.5}$$

But

$$\text{var}(W_{n+1} - I_n) = \text{var}(W_{n+1}) + \text{var}(I_n) - 2\,\text{cov}(W_{n+1} I_n). \tag{7.4.6}$$

Now $W_{n+1} I_n = 0$, and hence,

$$\begin{aligned}\text{cov}(W_{n+1} I_n) &= E(W_{n+1} I_n) - E(W_{n+1}) E(I_n) \\ &= -E(W)\left[\frac{1}{\lambda} - \frac{1}{\mu}\right].\end{aligned} \tag{7.4.7}$$

From Eqs. (7.4.5)–(7.4.7), we get

$$\sigma_W^2 + \sigma_u^2 + \sigma_v^2 = \sigma_W^2 + \text{var}(I_n) + 2E(W)\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)$$

so that

$$\text{var}(I_n) = \sigma_u^2 + \sigma_v^2 - 2\left(\frac{1}{\lambda} - \frac{1}{\mu}\right)E(W), \tag{7.4.8}$$

and putting the value of (7.4.8) in (7.4.4), we get

$$\text{var}(\tau_n) = \sigma_u^2 + 2\sigma_v^2 - \frac{2}{\lambda}(1-\rho)E(W). \tag{7.4.9}$$

Using the value of $E(W)$ from (7.2.3), we get

$$\begin{aligned}\text{var}(\tau_n) &= \sigma_u^2 + 2\sigma_v^2 - \left[\sigma_u^2 + \sigma_v^2 + \frac{1}{\lambda^2}(1-\rho)^2\right] + \frac{1}{\lambda}(1-\rho)\left(\frac{v_h^{(2)}}{v_h}\right) \\ &= \sigma_v^2 - \frac{(1-\rho)^2}{\lambda^2} + \left[\frac{(1-\rho)}{\lambda}\right]\left[\frac{v_h^{(2)}}{v_h}\right].\end{aligned} \tag{7.4.10}$$

■

7.4.1 Particular case

For an $M/G/1$ queue, $v_h^{(2)}/v_h = 2/\lambda$, so we get

$$\begin{aligned}\text{var}(\tau_n) &= \sigma_v^2 - \frac{(1-\rho)^2}{\lambda^2} + \frac{(1-\rho)}{\lambda}\frac{2}{\lambda} \\ &= \sigma_v^2 + \frac{1-\rho^2}{\lambda^2}.\end{aligned} \tag{7.4.11}$$

The variance of the output process of an $M/G/1$ system in steady state is exactly known when the mean and variance of the service-time distribution are exactly known. When the service time is exponential—that is, when the

system is $M/M/1$, $\sigma_v^2 = 1/\mu^2$ so that

$$\text{var}(\tau_n) = \frac{1}{\mu^2} + \frac{1-\rho^2}{\lambda^2} = \frac{1}{\lambda^2}.$$

As is well known, the output process of an $M/M/1$ queue in steady state is Poisson with the same rate as the arrival process.

7.4.2 Output process of a $G/G/c$ system

It is well known that the output process of $M/M/c$ and $M/G/\infty$ stationary queueing systems are Poisson. Whitt (1984b) examines the output process of a $G/G/c$ system. He shows that the output process in a large class of stationary $G/G/c$ systems is *approximately Poisson*, when there are many busy slow servers. Refer to Whitt for details and for limit theorems for the case when c and ρ increase.

7.5 Some Bounds for the $G/G/1$ System

7.5.1 Bound for $E(I)$

We have from (7.2.1) as $a_0 \leq 1$

$$E(I) \geq \frac{1-\rho}{\lambda} = \frac{1}{\lambda} - \frac{1}{\mu}, \tag{7.5.1}$$

which gives a lower bound for $E(I)$. The equality holds for the $D/D/1$ queue.

7.5.2 Bounds for $E(W)$

7.5.2.1 Upper bound

We have

$$v_h^{(2)} = E(I^2) = \text{var}(I) + [E(I)]^2 \geq E(I)^2$$

(from (7.5.1)) and so

$$\frac{v_h^{(2)}}{v_h} \geq E(I) \geq \frac{1}{\lambda}(1-\rho).$$

From (7.2.3), we get

$$\begin{aligned} E(W) &\leq \frac{\lambda^2(\sigma_u^2 + \sigma_v^2) + (1-\rho)^2}{2\lambda(1-\rho)} - \frac{1}{2\lambda}(1-\rho) \\ &= \frac{\lambda(\sigma_u^2 + \sigma_v^2)}{2(1-\rho)}, \end{aligned} \tag{7.5.2}$$

which is an upper bound for $E(W)$ for a $G/G/1$ system.

The equality holds for the $D/D/1$ queue. The importance of the bounds (7.5.1) and (7.5.2) is that they involve the first two moments of the arrival- and service-time distributions.

7.5.2.2 Lower bound

We shall discuss Marshall's lower bound as given next.

Theorem 7.4. *For a $G/G/1$ queue, $E(W) \geq r$, where r is the unique nonnegative root of the equation*

$$x = \int_{-x}^{\infty} \{1 - K(u)\}du; \tag{7.5.3}$$

the root is unique iff $\rho < 1$.

Proof: Let

$$f(x) = x - \int_{-x}^{\infty} \{1 - K(u)\}du;$$

we are to show that r is the unique nonnegative root of $f(x) = 0$. We have

$$\begin{aligned} f'(x) &= 1 - \{1 - K(-x)\} \\ &= K(-x) \geq 0 \quad \text{for} \quad x \geq 0, \end{aligned}$$

so that $f(x)$ is monotonically increasing for $x \geq 0$. For $x = 0$,

$$f(0) = 0 - \int_{0}^{\infty} \{1 - K(u)\}du < 0.$$

For large x, say $x \to A\,(>0)$, we have

$$\begin{aligned} f(A) &= A - \int_{-A}^{\infty} \{1 - K(u)\}du \\ &= A - \int_{-A}^{\infty} \left\{ \int_{u}^{\infty} dK(t) \right\} du \\ &= A - \int_{-A}^{\infty} \left\{ \int_{-A}^{t} du \right\} dK(t) \\ &\geq A - \int_{-\infty}^{\infty} (t + A)dK(t) \\ &= A - \left(\frac{1}{\mu} - \frac{1}{\lambda} \right) - A \\ &= \frac{1 - \rho}{\lambda} > 0 \quad \text{when} \quad \rho < 1. \end{aligned}$$

Thus, we have $f(0) < 0$ and $f(A) > 0$, where A is large and $\rho < 1$. Thus, $f(x) = 0$ has a unique nonnegative root when $\rho < 1$—that is, Eq. (7.5.3)

has a unique nonnegative root when $\rho < 1$. Denoting this root by r, we shall show that $E(W) \geq r$. Let

$$g(x) = \int_{-x}^{\infty} \{1 - K(u)\} du, \tag{7.5.4}$$

so that

$$f(x) = x - g(x).$$

As $f(0) < 0$, $f(r) = 0$, and $f(A) > 0$ for large A, we get

$$\begin{aligned} g(x) &> x \quad \text{when} \quad x < r \\ &\leq x \quad \text{when} \quad x \geq r. \end{aligned} \tag{7.5.5}$$

Thus, the function $g(x)$ is continuous and convex for $x \geq 0$. Again, since

$$W_{n+1} = \max(0,\ W_n + X_n),$$

we have that

$$Z = (W_{n+1} \mid W_n = x) = \max(0,\ X_n + x)$$

is positive, so that

$$E(Z) = E(W_{n+1} \mid W_n = x) = \int_0^{\infty} \{1 - G(t)\} dt,$$

where $G(t) = P(Z \leq t)$ is the DF of Z. Now

$$\begin{aligned} G(t) &= P\{Z \leq t\} \\ &= P\{0 \leq t\} P\{X_n + x \leq t\} \\ &= P\{X_n \leq t - x\} \\ &= K(t - x). \end{aligned}$$

Thus, for $x \geq 0$,

$$\begin{aligned} E(Z) &= \int_0^{\infty} \{1 - K(t - x)\} dt \\ &= \int_{-x}^{\infty} \{1 - K(v)\} dv, \quad \text{putting } v = t - x \\ &= g(x). \end{aligned} \tag{7.5.6}$$

Again using the relation

$$\begin{aligned} E(X) &= E\{E(X \mid Y = y)\} \\ &= \int_0^{\infty} E(X \mid Y = y) dF(y), \end{aligned} \tag{7.5.7}$$

where F is the DF of Y; we have, using (7.5.6),

$$\begin{aligned} E(W_{n+1}) &= E[E(W_{n+1} \mid W_n = x)] \\ &= \int_0^\infty E(W_{n+1} \mid W_n = x) dW(x) \\ &= \int_0^\infty g(x) dW(x), \end{aligned} \tag{7.5.8}$$

where $g(x)$ is a continuous convex function.

Using Jensen's inequality for the expected value of a convex function of a nonnegative RV, we get

$$E(W_{n+1}) \geq g[E(W_n)]. \tag{7.5.9}$$

In steady state, $E(W_n) = E(W_{n-1}) = E(W)$, so that

$$\begin{aligned} E(W) &\geq g(E(W)) \\ &= \int_{-E(W)}^\infty \{1 - K(u)\} du. \end{aligned} \tag{7.5.10}$$

Assume, if possible, that $E(W) < r$; then from (7.5.5)

$$g[E(W)] = \int_{-E(W)}^\infty \{1 - K(u)\} du > E(W) \quad \text{for } E(W) < r,$$

which contradicts (7.5.10). This contradiction is due to our assumption that $E(W) < r$. Thus, we must have

$$E(W) \geq r.$$

Finally, putting the upper and lower bounds together (from Eqs. (7.5.2) and (7.5.3)), we get

$$r \leq E(W) \leq \frac{\lambda(\sigma_u^2 + \sigma_v^2)}{2(1 - \rho)}. \tag{7.5.11}$$

■

Remarks:

(1) We can write Eq. (7.5.3) as

$$x = \alpha + \int_{-x}^0 \{1 - K(u)\} dx,$$

$$\text{where} \quad \alpha = \int_0^\infty \{1 - K(u)\} du. \tag{7.5.12}$$

Equation (7.5.3) has a solution *iff* the curves $y = x$ and

$$y = \alpha + \int_{-x}^{0} \{1 - K(u)\}du$$

intersect.

If $\alpha = 0$, then the second curve also passes through the origin, and $x = 0$ is a solution.

If $\alpha > 0$, then $y = \alpha > 0$ when $x = 0$, so that the second curve will lie above the first curve at the origin. Thus, the two curves will intersect if and only if, for x sufficiently large, the first curve will lie above the second curve, that is, *iff*

$$\begin{aligned} x &> \alpha + \int_{-x}^{0} \{1 - K(u)du \\ &= \alpha + x - \int_{-x}^{0} K(u)du, \end{aligned}$$

or *iff* $\int_{-x}^{0} K(u)du > \alpha = \int_{0}^{\infty}\{1 - K(u)\}du$, in other words, *iff*

$$\beta = \int_{-\infty}^{0} K(u)du > \int_{0}^{\infty}\{(1 - K(u)\}du = \alpha. \tag{7.5.13}$$

Now write

$$\begin{aligned} Z_1 &= \max(0,\ X_n) \\ Z_2 &= -\min(0,\ X_n) \\ F_1(t) &= Pr(Z_1 \le t) = Pr(0 \le t)\, Pr(X_n \le t) = K(t), \\ F_2(t) &= Pr(Z_2 \le t) = Pr\{-\min(0,\ X_n) \le t\} \\ &= Pr\{\min(0,\ X_n) \ge -t\} \\ &= Pr\{0 \ge -t\}\, Pr\{X_n \ge -t\} \\ &= 1 - K(-t). \end{aligned}$$

Since Z_1 and Z_2 are positive RVs,

$$E(Z_1) = \int_{0}^{\infty}\{1 - F_1(t)\}dt = \int_{0}^{\infty}\{1 - K(t)\}dt = \alpha \quad \text{and}$$

$$E(Z_2) = \int_{0}^{\infty}\{1 - F_2(t)\}dt = \int_{0}^{\infty} K(-t)dt = -\int_{-\infty}^{0} K(t)dt = -\beta,$$

so that (7.5.13) implies

$$E\{\min(0,\ X_n)\} > E\{\max(0,\ X_n)\}.$$

This leads to

$$\begin{aligned} E\{\max(0,\ X_n) - \min(0,\ X_n)\} < 0 \quad &\text{or} \\ E\{X_n\} < 0 \quad &\text{or} \\ E(v_n - u_n) < 0 \quad &\text{or} \\ \frac{1}{\mu} - \frac{1}{\lambda} < 0 \quad &\text{or} \quad \rho < 1. \end{aligned}$$

Thus, the two curves will intersect *iff* $\rho < 1$. The x-coordinate of the point of intersection will be the lower bound of the expected waiting time.

(2) Since $\sigma_u^2 + \sigma_v^2 > 0$ for all systems except for $D/D/1$, both the bounds tend to infinity if $\rho \to 1$.

(3) Rosberg (1987) derives new lower and upper bounds. The bounds obtained are functions of moments of order $r (r > 2)$ of u_n and v_n. It is claimed that these moments are better for low traffic intensity.

Example 7.4. The value of r for an $M/M/1$ queue is given by

$$r = \left(-\frac{1}{\lambda}\right) \log(1 - \rho^2).$$

Proof: For an $M/M/1$ queue

$$\begin{aligned} A(x) &= P(u_n \le x) = 1 - e^{-\lambda x} \quad \text{and} \\ B(x) &= P(v_n \le x) = 1 - e^{-\mu x}, \end{aligned}$$

so that from (7.1.4), for $x \ge 0$

$$\begin{aligned} K(x) &= \int_0^\infty \left\{1 - e^{-\mu(x+u)}\right\} \lambda e^{-\lambda u}\, du \\ &= 1 - \lambda \int_0^\infty e^{-(\lambda+\mu)u} e^{-\mu x}\, du. \end{aligned}$$

Thus,

$$K(x) = 1 - \frac{\lambda e^{-\mu x}}{\lambda + \mu}, \quad x \ge 0. \tag{7.5.14a}$$

For $x < 0$,

$$\begin{aligned} K(x) &= \int_0^\infty B(x + u) dA(u) \\ &= \int_0^\infty B(v) dA(v - x), \quad \text{since } x < 0 \\ &= \lambda \left[\int_0^\infty e^{-(v-x)\lambda}\, dv - \int_0^\infty e^{-(\mu+\lambda)v} e^{x\lambda}\, dv \right]. \end{aligned}$$

Thus,

$$K(x) = e^{\lambda x} - \frac{\lambda e^{\lambda x}}{\lambda + \mu} = \frac{\mu e^{\lambda x}}{\lambda + \mu}, \quad x < 0. \tag{7.5.14b}$$

The lower bound r is the unique positive root of

$$\begin{aligned}
0 &= x - \int_{-x}^{\infty} \{1 - K(t)\} dt \\
&= x - \int_{-x}^{0} \{1 - K(t)\} dt - \int_{0}^{\infty} \{1 - K(t)\} dt \\
&= x - \int_{-x}^{0} \left\{1 - \frac{\mu e^{\lambda t}}{\lambda + \mu}\right\} dt - \int_{0}^{\infty} \frac{\lambda}{\lambda + \mu} e^{-\mu t} dt \\
&= x - x + \frac{\mu}{\lambda(\lambda + \mu)} (1 - e^{-\lambda x}) - \frac{\lambda}{\lambda + \mu} \frac{1}{\mu} \\
&= \frac{\mu^2 - \lambda^2 - \mu^2 e^{-\lambda x}}{\lambda\mu(\lambda + \mu)} \\
&= \frac{\mu - \lambda}{\lambda\mu} - \frac{\mu e^{-\lambda x}}{\lambda(\lambda + \mu)},
\end{aligned}$$

whence

$$e^{-\lambda x} = (1 - \rho^2)$$

so that the root r is given by

$$r = -\frac{1}{\lambda} \log_e(1 - \rho^2) = -\frac{1}{\lambda} \ln(1 - \rho^2). \tag{7.5.15}$$

■

Note: The lower bound tends to ∞ as $\rho \to 1$. The upper bound is

$$\frac{\lambda\left(\frac{1}{\lambda^2} + \frac{1}{\mu^2}\right)}{2\left(1 - \frac{\lambda}{\mu}\right)} = \frac{1}{\lambda} \frac{1 + \rho^2}{2(1 - \rho)},$$

which also tends to ∞ as $\rho \to 1$. Thus, as $\rho \to 1$, both the bounds tend to ∞. As $\rho \to 0$, the lower bound tends to 0 and the upper bound tends to $1/2\lambda$.

The upper and lower bounds and the true value of $E(W)$ for fixed $\lambda = 1$ and varying μ—that is, for $\rho = 0$ to $\rho = 1$—can be seen in Marshall (1968b). Another lower bound has been put forward by Marchal (1978), which we discuss shortly.

Theorem 7.5. *For a G/G/1 queue*

$$E(W) \geq \frac{\lambda^2\sigma_v^2 + \rho(\rho - 2)}{2\lambda(1 - \rho)}.$$

Marchal derives the preceding lower bound from the lower bound obtained by Kingman (1962), which is as follows:

$$E\{W\} \geq \frac{\lambda}{2(1 - \rho)} E\{[\max(0, v - u)]^2\}. \tag{7.5.16}$$

For details of proof and other results, such as bounds of $G/G/c$ and of $M/G/c$ systems, refer to Marchal (1978).

See also Kingman (1970), Schassberger (1970), Ross (1974), Stoyan (1977), and Sobel (1980).

Notes:

(1) The lower bound depends on the mean of u and the first two moments of v.

(2) Marchal's lower bound is positive *iff*

$$\sigma_v^2 \geq \frac{2 - \rho}{\lambda\mu}.$$

Thus, this lower bound is of use only when this inequality is satisfied.

Example 7.5. The bounds of $\text{var}(\tau_n)$ are given by

$$\sigma_v^2 \leq \text{var}(\tau_n) \leq \sigma_u^2 + 2\sigma_v^2 - 2r\left(\frac{1}{\lambda} - \frac{1}{\mu}\right).$$

We have from (7.4.9)

$$\text{var}(\tau_n) = \sigma_u^2 + 2\sigma_v^2 - \left(\frac{2}{\lambda}\right)(1 - \rho)E(W).$$

Using (7.5.11), we get at once

$$\sigma_v^2 \leq \text{var}(\tau_n) \leq \sigma_u^2 + 2\sigma_v^2 - 2r\frac{(1 - \rho)}{\lambda}.$$

Remarks:

(1) It may be noticed that the moments of the idle-time distribution figured in many of the expressions obtained. The idle-time distribution is some complicated tail distribution of interarrival time. Marshall (1968a) observed that

by placing some restrictions on the interarrival-time distribution, it is possible to obtain some desirable properties of the moments of the idle-time distribution. Marshall considered some special type of interarrival-time distribution (having bounded mean residual life) and obtained bounds for $E(I^2)/2E(I)$. In another paper (1968b) he considered further generalizations. (See Problems and Complements.)

(2) Another approach to a $G/G/1$ queue through Wiener-Hopf factorization is considered in detail by Prabhu (1980): ladder processes and the random walk $\{S_n\}$, where

$$S_n = X_1 + \cdots + X_n,$$

prove very important in this approach. Denote

$$\begin{aligned} \bar{N} &= \min\{n > 0;\ S_n \leq 0\} \quad \text{and} \\ g_n(x) &= Pr\{\bar{N} = n,\ \ S_{\bar{N}} \leq x\} \\ &= Pr\{S_1 > 0,\ S_2 > 0, \ldots,\ S_{n-1} > 0,\ \ S_n \leq x < 0\}. \end{aligned}$$

Then $\bar{N}$ is the number served during a busy period, and $I = -S_{\bar{N}}$ is the idle period. The distribution of $\{W_n\}$ is completely solved, and the joint distribution of $\bar{N}$ and I has been obtained in explicit form for $M/M/1$ and $G/M/1$ systems. (See Problems and Complements.) Ladder processes, which were primarily used for the study of waiting times and idle periods of a $G/G/1$ queue, have also been found to be important in the study of queue-length processes in some special cases. (See Prabhu (1980) for further results for the $G/G/1$ queue; see also Whitt (1980, 1984a).)

Problems and Complements

7.1. Using (7.5.14a) and (7.5.14b), show that the (two-sided) LST of the RV $X_n = v_n - u_n$ for an $M/M/1$ system is given by

$$K^*(s) = \frac{\lambda\mu}{(\lambda - s)(\mu + s)}.$$

Verify that $K^*(s) = B^*(s)\,A^*(-s)$.

7.2. Show that the DF $K(x) = P\{X_n \leq x\}$, $X_n = v_n - u_n$ for the system $G/M/1$ is given by

$$\begin{aligned} K(x) &= \int_{-x}^{\infty} dA(u)\left\{1 - e^{-\lambda(u+x)}\right\}, \quad x \leq 0, \quad \text{and} \\ &= 1 - ce^{-\lambda x}, \quad x \geq 0, \end{aligned}$$

where $c = \psi(\theta) = \int_0^\infty e^{-\theta x}\, dA(x)$ is the LST of the interarrival-time distribution. Further show that $K(x)$ has the partial lack of memory

$$P\{X_n \leq y \mid X_n > x\} = 1 - e^{-\lambda(y-x)}, \quad y > x \geq 0.$$

(Prabhu, 1980)

7.3. For an $M/D/1$ system, find $K^*(s)$, $K(u)$; find $E(X_n)$ from $K^*(s)$.

7.4. Obtain the upper and lower bounds of $E(W)$ for a $D/M/1$ queue.

7.5. Using Eq. (7.1.25), find the waiting-time distribution for the $D/D/1$ queue.

7.6. For the sample sequence $\{W_n, n \geq 0\}$ with first eight elements $A = \{W_0 = 0, W_1 > 0, W_2 = 0, W_3 > 0, W_4 > 0, W_5 = 0, W_6 > 0, W_7 > 0\}$ compute the W_ns in terms of S_ns, and the I_ns in terms of S_ns, and the number served during the busy periods corresponding to A (with notations as in Section 7.1).

7.7. Show that

$$W_n = \max(0, X_n + X_{n-1} + \cdots + X_{n-r+1}, 1 \leq r \leq n).$$

7.8. Find the expressions for the mean and the variance of the waiting time for the systems (i) $E_2/M/1$ and (ii) $M/G/1$.

7.9. Batch service with fixed batch size k. Let $W_{k,n}$ be the waiting time in queue of the last (kth) person who arrived and formed the nth batch of k customers. Then show that the expected waiting time for an unspecified customer $E(W)$ is given by

$$E(W) = E(W_{k,\cdot}) + \frac{(k-1)}{2\lambda}.$$

Deduce that Little's formula is satisfied for $M/G^k/1$ with services in batches of size k (Marshall, 1968b).

7.10. Suppose that there is a random start-up time R after every idle period so that the first customer in every busy period suffers a random delay R before his service commences, and assume that the idle time WRT the server is $I + R$. Then

$$E(W) = \frac{E(U^2)}{-2E(U)} + \frac{E(R^2) - E(I^2)}{2[E(R) + E(I)]} + \frac{\operatorname{cov}(W_n, U_n)}{E(U)},$$

where $U_n = v_n - u_n$; and assuming independence of I and R, we get that $E(W)$ is given by the first two terms on the RHS (Marshall, 1968b).

Note: A specific example of the preceding class of queues is the single-service queue with N-policy, where the server starts his first service in a busy period only after N customers arrive and then serves one by one until all the customers are served. See Sections (6.4.5) and (8.3.3) for a queue under N-policy.

7.11. Marchal's weighting factors.
As an approximation for $E(W)$, Marchal suggests some weighting factors to be used with the upper bound. The weighting factor

$$\frac{\rho^2 + \lambda^2\sigma_v^2}{1 + \lambda^2\sigma_v^2}$$

(which tends to 1 as ρ tends to 1) used to scale down the upper bound leads to the approximation

$$E(W) \simeq \frac{\lambda(\sigma_u^2 + \sigma_v^2)}{2(1-\rho)} \frac{\rho^2 + \lambda^2\sigma_v^2}{1 + \lambda^2\sigma_v^2}.$$

When the arrival process is Poisson, show that this leads to the Pollaczek-Khinchin formula—that is, it is exact for $M/G/1$. Show that it works well with a $G/M/1$ system also.

Another weighting factor suggested is

$$\frac{\rho^2\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2}$$

(which also tends to 1 as ρ tends to 1). Using this to scale down the upper bound, show that one can get as an approximation

$$E(W) \simeq \frac{\rho(\lambda^2\sigma_u^2 + \mu^2\sigma_v^2)}{2\mu(1-\rho)}.$$

Show that it is exact for $M/G/1$ and $D/D/1$ (Marchal, 1978).

7.12. Distribution with bounded mean residual life. A nondiscrete distribution F is said to have its mean residual life bounded above (below) by γ, denoted by γ-MRLA (γ-MRLB) *iff*

$$\int_t^\infty \frac{\{1 - F(u)\}du}{\{1 - F(t)\}} \underset{(\geq)}{\leq} \gamma \quad \text{for all } t \geq 0,\ \gamma < \infty.$$

For a $G/G/1$ queue where the interarrival-time distribution is γ-MRLA (γ-MRLB), show that

$$\frac{E(I^2)}{2E(I)} = \frac{v_h^{(2)}}{2v_h} \underset{(\geq)}{\leq} \gamma.$$

(Marshall, 1968a)

7.13. Show that for an $M/M/1$ queue

$$E\{z^{\tilde{N}}e^{itI}\} = \frac{\lambda\zeta(z)}{\lambda - it},$$

where

$$\zeta(z) = \frac{[(\lambda+\mu) - \sqrt{\{(\lambda+\mu)^2 - 4\lambda\mu z\}}]}{2\lambda}.$$

Write down the PGF of $\tilde{N}$ and the PDF of I (Prabhu, 1980).

7.14. Show that for a $G/M/1$ queue

$$E\{z^{\tilde{N}}e^{-itI}\} = \frac{\mu\xi - \mu z\phi_1(-t)}{\mu\xi - \mu + it}, \quad 0 < z < 1, \quad t \text{ real},$$

where ϕ_1 is the characteristic function of the RV having DF $A(u) = Pr\{u_n \leq u\}$ and $\xi = \xi(z)$ is the unique continuous solution of the equation

$$\xi = z\psi(\mu - \mu\xi)$$

in the interval $0 < z < 1$, $\psi(\theta)$ being the LST of the RV having DF $A(u)$ (Prabhu, 1980).

References and Further Reading

Björklund, M., and Elldin, A. (1964). A practical method for calculation of certain types of complex common control systems. *Ericsson Technics* **20**, 3–75.

Boxma, O. J., Cohen, J. W., and Huffels, N. (1979). Approximations of the mean waiting time in an $M/G/s$ queueing system. *Opns. Res.* **27**, 1115–1127.

Brumelle, S. L. (1973). Bounds on the wait in a $GI/M/k$ queue. *Mgmt. Sci.* **19**, 773–777.

Cohen, J. W. (1982). *The Single Server Queue*, 2nd ed., North-Holland, Amsterdam.

Cosmetatos, G. P. (1975). Approximate explicit formulae for the average queueing time in the processes $(M/D/r)$ and $(D/M/r)$. *INFOR*, **13**, 328–332.

Cosmetatos, G. P. (1976). Some approximate equilibrium results for the multiserver queue $(M/G/r)$. *Oper. Res. Qrly.* **27**, 615–620.

Fredericks, A. A. (1982). A class of approximations for the waiting time distribution in a $GI/G/1$ queue. *Bell Syst. Tech. J.* **61**, 295–325.

Gross, D., and Harris, C. M. (1985). *Fundamentals of Queueing Theory*, 2nd ed., Wiley, New York.

Harel, A., and Zipkin, P. (1987). Strong convexity results for queueing systems. *Opns. Res.* **35**, 405–418.

Heyman, D. P. (1968). Optimal control policies for $M/G/1$ queueing systems. *Opns. Res.* **16**, 362–382.

Hillier, F. S., and Yu, O. S. (1981). *Queueing Tables and Graphs*. North-Holland, New York.

Hokstad, P. (1978). Approximations for the $M/G/m$ queue. *Opns. Res.* **26**, 510–523.

Jagerman, D. (1987). Approximations for waiting time in $GI/G/1$ systems. *Queueing Systems* **2**, 351–362.

Jain, J. L., and Grassman, W. K. (1989). Numerical solution for the departure process from $GI/G/1$ queue. *Comp. & Opns. Res.* **15**, 293– 296.

Kimura, T. (1986). A two-moment approximation for the mean waiting time in $GI/G/s$ queue. *Mgmt. Sci.* **32**, 751–763.

Kimura, T. (1991a). Refining Cosmetatos' approximation for the mean waiting time in the $M/D/s$ queue. *J. Opnl. Res. Soc.* **42**, 595–603.

Kimura, T. (1991b). Approximating the mean waiting time in $GI/G/s$ queue. *J. Oper. Res. Soc.* **42**, 959–970.

Kimura, T. (1993). Approximation for the Delay Probability in the $M/G/1$ Queue in *Stochastic Models in Engineering, Technology and Management* (Eds. S. Osaki and D. N. P. Murthy), 277–285, World Scientific.

Kimura, T. (1994). Approximations for multi-server queues: system interpolations. *Queueing Systems* **17**, 347–382. (Contains a list of 82 references.)

Kingman, J. F. C. (1962). Some inequalities for the queue $GI/G/1$. *Biometrika* **49**, 315–324.

Kingman, J. F. C. (1970). Inequalities in the theory of queues. *J.R.S.S.* **B32**, 102–110.

Kleinrock, L. (1976). *Queueing Systems, Vol. II: Computer Applications*, Wiley, New York.

Krämer, W., and Langenbach-Belz, M. (1988). Approximate formulae for the delay in the queueing system $GI/G/1$, *Proc. 8th Intl. Teletraffic Congress*, Melbourne 1–8.

Lee, A. M., and Longton, P. A. (1957). Queueing process associated with airlines passenger check-in, *Oper. Res. Qrly.* **10**, 56–71.

Lindley, D. V. (1952). The theory of queues with a single server. *Proc. Camb. Phil. Soc.* **48**, 277–289.

Maaløe, E. (1973). Approximation formulae for estimation of waiting-time in multiple-channel queueing systems. *Mgmt. Sci.* **19**, 703–710.

Marchal, W. G. (1978). Some simpler bounds on the mean queueing time. *Opns. Res.* **26**, 1083–1088.

Marshall, K. T. (1968a). Some inequalities in queueing. *Opns. Res.* **16**, 651–665; comments (by R. V. Evans), 666–668.

Marshall, K. T. (1968b). Bounds for some generalizations of the $GI/G/1$ queue. *Opns. Res.* **16**, 841–848.

Marshall, K. T. (1968c). Some relationships between the distributions of waiting time, and inter-output time in the $GI/G/1$ queue. *SIAM J. Appl. Math.* **16**, 324–327.

Newell, G. F. (1971). *Applications of Queueing Theory* (2nd Ed., 1982), Chapman and Hall, London.

Nozaki, S. A., and Ross, S. M. (1978). Approximations for the queue-length distribution of an $M/GI/s$ queue by the basic equations. *J. Appl. Prob.* **15**, 826–894.

Page, E. (1972). *Queueing Theory in OR*, Butterworth, London.

Page, E. (1982). Tables of waiting times for $M/M/n$, $M/D/n$ and $D/M/n$ and their use to give approximate waiting times in more general queues. *J. Opnl. Res. Soc.* **33**, 453–473.

Prabhu, N. U. (1980). *Stochastic Storage Processes*, Springer-Verlag, New York.

Rosberg, Z. (1987). Bounds on the expected waiting time in a $GI/G/1$ queue: upgrading for lower traffic intensity. *J. Appl. Prob.* **24**, 749–757.

Ross, S. M. (1974). Bounds on the delay distribution in $GI/G/1$. *J. Appl. Prob.* **24**, 749–757.

Schassberger, R. (1970). On the waiting time in queueing system $GI/G/1$. *Ann. Math. Statist.* **41**, 182–187.

Seelen, L. P., Tijms, H. C., and van Hoorn, M. H. (1985). *Tables for Multi-Server Queues*, North Holland, Amsterdam.

Shore, H. (1988a). Simple approximations for the $GI/G/C$ queue—I: The steady state probabilities. *J. Oper. Res. Soc.* **39**, 279–284.

Shore, H. (1988b). Simple approximations for the $GI/G/C$ queue—II: The moments, the inverse distribution function and the loss function of the number in the system and of the queue delay. *J. Oper. Res. Soc.* **39**, 381–391.

Sobel, M. (1980). Simple inequalities for multiserver queues. *Mgmt. Sci.* **26**, 951–956.
Stoyan, D. (1976). Approximations for $M/G/s$ queues. *Math. Oper. Statist.* **7**, 587–594.
Stoyan, D. (1977). Bounds and approximations in queueing through monotonicity and continuity. *Opns. Res.* **25**, 851–863.
Tijms, H. C. (1987). A Quick and Practical Approximation to the Waiting Time Distribution in the Multi-Server Queue with Priorities in *Computer Performance and Reliability,* pp. 161–169 (Eds. Tazeolla *et al.*), North Holland, Amsterdam.
Tijms, H. C. (1994). *Stochastic Models: An Algorithmic Approach,* Wiley, Chichester.
Tijms, H. C., Van Hoorn, M. H., and Federgruen, A. (1981). Approximations for the steady-state probabilities in the $M/G/c$ queue, *Adv. Appl. Prob.* **13**, 186–206.
Van Hoorn, M. H., and Tijms, H. C. (1982). Approximations for the waiting time distribution of the $M/G/c$ queue. *Perf. Eval.* **2**, 22–28.
Weber, R. R. (1983). A note on the waiting times in single server queues. *Opns. Res.* **31**, 950–951.
Whitt, W. (1980). The effect of variability in $GI/G/s$ queue. *J. Appl. Prob.* **17**, 1062–1071.
Whitt, W. (1984a). Minimizing delays in the $GI/G/1$ queue. *Opns. Res.* **32**, 41–51.
Whitt, W. (1984b). Departures from a queue with many servers. *Math. Opns. Res.* **9**, 534–544.
Whitt, W. (1989). An interpolation approximation for the mean workload in the $GI/G/1$ queue. *Oper. Res.* **37**, 936–952.

This Page Intentionally Left Blank

Miscellaneous Topics

8.1 Heavy-Traffic Approximation for Waiting-Time Distribution

8.1.1 Kingman's heavy-traffic approximation for a $G/G/1$ queue

A queue with traffic intensity barely less than unity is called a heavy-traffic queue. The behavior of a queueing system $G/G/1$ in the heavy-traffic case was first investigated by Kingman (1961). Here we discuss Kingman's result, which is a sort of Central Limit Theorem for heavy traffic and is given next.

Theorem 8.1. *Under heavy traffic, the steady-state waiting-time distribution in a $G/G/1$ queue can be approximated by an exponential distribution.*

Proof: The starting point in this study is Eq. (7.1.13) of Chapter 7.

$$\bar{W}(s) = \frac{\bar{W}^{-}(s)}{B^*(s)A^*(-s) - 1},$$

where $\bar{W}(s)$ is the LT of the waiting time W, while $B^*(s)$ and $A^*(s)$ are the LSTs of the service-time and interarrival-time distributions v and u, respectively, and $\bar{W}^{-}(s)$ is the two-sided LT of W^{-} defined in Section 7.1.2. The preceding is written as

$$B^*(s)A^*(-s) - 1 = \frac{\bar{W}^{-}(s)}{\bar{W}(s)}. \tag{8.1.1}$$

Maclaurin's expansion of $B^*(s)$ gives

$$B^*(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} B^{*(k)}(0), \tag{8.1.2}$$

where $B^{*(k)}(0)$ is the kth derivative of $B^*(s)$ at $s = 0$. Denote $E(v^k) = b_k$, then

$$B^{*(k)}(0) = (-1)^k E(v^k) = (-1)^k b_k.$$

Truncating the series on the RHS of (8.1.2) to three terms, we get

$$B^*(s) = 1 - b_1 s + \frac{b_2}{2} s^2 + O(s^2). \tag{8.1.3}$$

Similarly, writing $E(u^k) = a_k$, and truncating to three terms, we get

$$A^*(-s) = 1 + a_1 s + \frac{a_2}{2} s^2 + O(s^2). \tag{8.1.4}$$

Thus,

$$\begin{aligned} B^*(s)A^*(-s) - 1 &= \left(1 - b_1 s + \frac{b_2}{2} s^2\right)\left(1 + a_1 s + \frac{a_2 s^2}{2}\right) - 1 + O(s^2) \\ &= s\left[(a_1 - b_1) + \left(\frac{a_2}{2} + \frac{b_2}{2} - a_1 b_1\right)s\right] + O(s^2). \end{aligned} \tag{8.1.5}$$

Clearly $s = 0$ is one of the roots of the LHS of (8.1.5). There is another root s_0 that satisfies

$$(a_1 - b_1) + \left(\frac{a_2}{2} + \frac{b_2}{2} - a_1 b_1\right)s_0 = 0. \tag{8.1.6}$$

Denote

$$a_1 = \frac{1}{\lambda}, \qquad a_2 - a_1^2 = \sigma_u^2, \qquad b_1 = \frac{1}{\mu}, \qquad b_2 - b_1^2 = \sigma_v^2.$$

Then

$$a_1 - b_1 = \left(\frac{1}{\lambda}\right)(1 - \rho), \qquad \rho = \frac{\lambda}{\mu} \quad \text{and}$$

$$\begin{aligned} \frac{a_2}{2} + \frac{b_2}{2} - a_1 b_1 &= \frac{1}{2}\left(\sigma_u^2 + \frac{1}{\lambda^2}\right) + \frac{1}{2}\left(\sigma_v^2 + \frac{1}{\mu^2}\right) - \frac{1}{(\lambda\mu)} \\ &= \frac{1}{2}(\sigma_u^2 + \sigma_v^2) + \left(\frac{1}{2\lambda^2}\right)(1 - \rho)^2. \end{aligned}$$

Under the heavy-traffic condition $\rho \simeq 1$, and thus, the second term of the preceding can be neglected, so that from (8.1.6), we get

$$s_0 = -\frac{\left(\frac{1}{\lambda}\right)(1-\rho)}{\left\{\frac{1}{2}\left(\sigma_u^2 + \sigma_v^2\right)\right\}} = -\frac{2(1-\rho)}{\lambda\left(\sigma_u^2 + \sigma_v^2\right)}. \tag{8.1.7}$$

Thus, we can rewrite (8.1.5), as an approximation near $s = 0$, as

$$B^*(s)\,A^*(-s) - 1 = s(s - s_0)\left\{\frac{\left(\sigma_u^2 + \sigma_v^2\right)}{2}\right\} = s(s - s_0)K, \tag{8.1.8}$$

where

$$K = \frac{1}{2}\left(\sigma_u^2 + \sigma_v^2\right).$$

From (8.1.1), we have, near the origin,

$$s(s - s_0)K = \frac{\bar{W}^-(s)}{\bar{W}(s)} \quad \text{or} \tag{8.1.9}$$

$$\begin{aligned} \bar{W}^-(s) &= s(s - s_0)K\,\bar{W}(s) \\ &= (s - s_0)K[s\,\bar{W}(s)] \\ &= (s - s_0)K\,W^*(s), \end{aligned} \tag{8.1.10}$$

where $W^*(s)$ is the LST of W—that is,

$$\begin{aligned} W^*(s) &= \int_0^\infty e^{-st}\,dW(t) = s\int_0^\infty e^{-st}W(t)dt \\ &= s\,\bar{W}(s). \end{aligned}$$

Since near the origin $W^*(s) = 1$, we get

$$\bar{W}^-(s) = -s_0 K.1.$$

Thus, from (8.1.1) and (8.1.8), we have

$$\begin{aligned} \bar{W}(s) &= \frac{-s_0 K}{s(s - s_0)K} = -\frac{s_0}{s(s - s_0)} \\ &= \frac{1}{s} - \frac{1}{s - s_0}. \end{aligned} \tag{8.1.11}$$

Inverting the LT, we get

$$W(t) = 1 - \exp(s_0 t)$$
$$= 1 - \exp\left(-\frac{2(1-\rho)}{\lambda(\sigma_u^2 + \sigma_v^2)} t\right), \tag{8.1.12}$$

which gives the distribution function of the waiting time to an approximation. The distribution is exponential with mean given by

$$E(W) \simeq \frac{\lambda(\sigma_u^2 + \sigma_v^2)}{2(1-\rho)}, \tag{8.1.13a}$$

that is, with parameter $1/E(W) = -s_0$.

The result (8.1.13a) for large $\rho(<1)$ can also be expressed as

$$E\,W(G/G/1) \simeq \frac{c_u^2 + c_v^2}{2} E\,W(M/M/1), \tag{8.1.13b}$$

where $c_u^2(c_v^2)$ is the square of the coefficient of variation of interarrival (service) time. ■

Remarks:

(1) The result showing the exponential character of the waiting-time distribution of a $G/G/1$ queue may be called the *Central Limit Theorem* for queueing theory. The result is extremely robust.

(2) In Section 7.5 we discussed the upper bound for the average waiting time $E(W)$ in a $G/G/1$ system for $0 \leq \rho \leq 1$ (as obtained by Kingman (1962b)). He shows that

$$E(W) \leq \frac{\sigma_u^2 + \sigma_v^2}{2\left(\frac{1}{\lambda}\right)(1-\rho)}. \tag{8.1.13}$$

This result gives an upper bound of $E(W)$ for $0 \leq \rho \leq 1$. The bound is reached when $\rho \to 1$; this result has been obtained in this section. This shows that heavy-traffic mean waiting time forms the strict upper bound for the mean waiting time in a $G/G/1$ system.

(3) The distribution function (8.1.12) is of the form

$$W(t) = 1 + ke^{s_0 t}$$
$$= 1 + s_0 E(W) e^{s_0 t}. \tag{8.1.12a}$$

The constant

$$k = s_0 E(W) \tag{8.1.12b}$$

equals -1 in the heavy-traffic situation, as previously discussed.

Fredericks (1982) proposes a class of approximations for a $G/G/1$ waiting-time-distribution function of the form (8.1.12a) and develops a procedure to estimate the parameters k and s_0. The dominant root and the mean delay, whenever known, can be used to determine the (constant) coefficient k, by using (8.1.12b).

Particular case. For an $M/G/1$ queue

$$A^*(s) = \frac{\lambda}{\lambda + s}$$

so that $A^*(-s)B^*(s) - 1 = 0$ reduces to

$$s\left[(1 - \lambda b_1) + \left(\frac{\lambda b_2}{2}\right)s\right] = 0.$$

Thus,

$$s_0 = -\frac{2(1-\rho)}{(\lambda b_2)}$$

and for exponential service time $b_2 = 2/\mu^2$, so that the mean waiting time $E(W)$ for an $M/M/1$ is given by

$$E(W) = -\frac{1}{s_0} = \frac{\rho}{\mu(1-\rho)}.$$

Notes: Kingman's result holds for very heavy traffic for the general $G/G/1$ queue. The waiting-time distribution even in $M/G/1$ queues for arbitrary ρ is somewhat difficult to compute. Benes (1956) has shown that the distribution can be expressed as a geometrically decreasing weighted sum of convolutions of residual service times. The approximation is convenient for light traffic.

Marchal (1987) has given an empirical extension of Kingman's result for the Poisson input queue. Marchal extends the heavy-traffic result of Kingman to conditions of moderate traffic intensity. We now discuss the result.

8.1.2 Empirical extension of the $M/G/1$ heavy-traffic approximation

Assume that the input is Poisson, so that

$$A^*(s) = \frac{\lambda}{\lambda + s}.$$

Taking two more terms of Maclaurin's expansion, we get

$$B^*(s) = 1 - b_1 s + \frac{b_2}{2}s^2 - \frac{b_3}{6}s^3 + \frac{b_4}{24}s^4 + O(s^4). \qquad (8.1.14)$$

Then the characteristic equation $A^*(-s)B^*(s) = 1$ reduces to

$$\left[\frac{\lambda}{(\lambda - s)}\right]\left[1 - b_1 s + \frac{b_2}{2}s^2 - \frac{b_3}{6}s^3 + \frac{b_4}{24}s^4\right] = 1$$

when all the terms above the fourth degree are ignored. The preceding can be written as $sf(s) = 0$, where

$$f(s) \equiv \left[(1 - \lambda b_1) + \left(\frac{\lambda b_2}{2}\right)s - \left(\frac{\lambda b_3}{6}\right)s^2 + \left(\frac{\lambda b_4}{24}\right)s^3\right] = 0 \quad (8.1.15)$$

is a cubic in s.

Now $f(-s)$ has only one variation of sign and, thus, applying Descartes's rule of signs, one finds that $f(s)$ possesses exactly one negative root, which we denote by s_0.

We consider Fredericks's approximation of the waiting-time distribution for a $G/G/1$ queue

$$W(t) = 1 + ke^{s_0 t}. \quad (8.1.16)$$

The mean $E(W)$ for an $M/G/1$ queue is known from the Pollaczek-Khinchin formula

$$E(W) = \frac{\lambda}{2(1 - \rho)}\left(\frac{1}{\mu^2} + \sigma_v^2\right). \quad (8.1.17)$$

The constant k can be estimated from

$$k = s_0 E(W).$$

Thus, we get an estimate of $W(t)$. The preceding approximation has the following advantages.

Maclaurin expansion of $B^*(s)$, which involves four moments of the service-time distribution, gives a better approximation than the one involving two moments. The result has less dependence on the asymptotic effect of the heavy traffic (i.e., of traffic intensity near unity). The cubic equation has exactly one negative root, s_0. Whenever $E(W)$ is known or can be estimated, one can easily estimate k. For the $G/G/1$ queue, one can use Marchal's (1976) estimate

$$E(W) = \frac{\lambda(\sigma_u^2 + \sigma_v^2)}{2(1 - \rho)} \cdot \frac{\rho^2 + \lambda^2\sigma_v^2}{1 + \lambda^2\sigma_v^2}. \quad (8.1.18)$$

The arguments of Marchal can be further extended to the case of a $G/G/1$ queue by starting with Maclaurin's expansion of $A^*(s)$ and $B^*(s)$.

Remarks:

(1) Marchal has considered numerical examples for same particular $M/G/1$ queue, with $G \equiv M$, $G \equiv D$, and $G \equiv E_2$. The approximations appear to be satisfactory for moderate traffic intensity, say, in the range 0.50–0.80.

(2) It can be verified that $k \to -1$ as $\rho \to 1$.

8.1.3 $G/M/c$ queue in heavy traffic

Consider a $G/M/c$ queue in steady state for which the conditional waiting time, given that an arrival has to wait, is exponential with mean $1/c\mu(1-\sigma)$ where σ is the root of the equation

$$\sigma = A^*(c\mu - c\mu\sigma), \tag{8.1.19}$$

$A^*(.)$ being the LST of the interarrival-time distribution. In heavy traffic where an arrival has to wait, the unconditional distribution of the waiting time approaches the conditional distribution. Thus, for $\rho \to 1$ the mean waiting time is given by

$$E(W) \simeq \frac{1}{c\mu(1-\sigma)} \tag{8.1.20}$$

Change the variable in (8.1.19) using

$$\alpha = c\mu(1-\sigma); \tag{8.1.21}$$

then (8.1.19) becomes

$$\sigma = A^*(\alpha) = 1 - \frac{\alpha}{c\mu}. \tag{8.1.22}$$

Expanding $A^*(\alpha)$ as a power series in α, we get

$$\begin{aligned} A^*(\alpha) &= 1 + \alpha\left\{\frac{d}{d\alpha}A^*(\alpha)|_{\alpha=0}\right\} \\ &\quad + \frac{\alpha^2}{2}\left\{\frac{d^2}{d\alpha^2}A^*(\alpha)|_{x=0}\right\} + O(\alpha^2) \\ &= 1 - \alpha E(u) + \frac{\alpha^2}{2}E(u^2) + O(\alpha^2). \end{aligned} \tag{8.1.23}$$

For heavy traffic, $E(W)$ is large; from (8.1.20) we get $\sigma \simeq 1$ and α small so that $O(\alpha^2)$ may be neglected. Thus, from (8.1.22) and (8.1.23), we get

$$1 - \frac{\alpha}{c\mu} = 1 - \frac{\alpha}{\lambda} + \frac{\alpha^2}{2}E(u^2).$$

Thus, we have

$$\alpha = \frac{2\left(\frac{1-\rho}{\lambda}\right)}{E(u^2)} = \frac{2\left(\frac{1}{\lambda}\right)(1-\rho)}{\sigma_u^2 + \{E(u)\}^2}. \tag{8.1.24}$$

For heavy traffic $\rho \to 1$, that is,

$$\frac{E(v)}{cE(u)} \simeq 1;$$

for exponential service time

$$[E(u)]^2 = \frac{\sigma_v^2}{c^2}$$

and so

$$\alpha = \frac{2\left(\frac{1}{\lambda}\right)(1-\rho)}{\sigma_u^2 + \frac{\sigma_v^2}{c^2}}. \tag{8.1.25}$$

Thus, the approximate value of the mean waiting time in a $G/M/c$ queue in heavy traffic is obtained as

$$E(W) \simeq \frac{1}{\alpha} = \frac{\sigma_u^2 + \frac{\sigma_v^2}{c^2}}{2\left(\frac{1}{\lambda}\right)(1-\rho)}. \tag{8.1.26}$$

Remarks:

(1) The preceding result for a $G/M/c$ system led Kingman (1964) to make the conjecture that for heavy traffic, the waiting time for a $G/G/c$ queue should be exponentially distributed with mean equal to that given by (8.1.26). Köllerström (1974) has proved the conjecture; the approximate distribution $W(t)$ of the waiting time in a $G/G/c$ queue in heavy traffic is exponential and is given by

$$W(t) \simeq 1 - \exp\left\{-\frac{2\frac{(1-\rho)}{\lambda}}{\sigma_u^2 + \frac{\sigma_v^2}{c^2}}t\right\}. \tag{8.1.27}$$

(2) Kingman (1961) was perhaps the first to study the asymptotic behavior of queues. His work, in a sense, motivated subsequent studies on asymptotic behavior, including the diffusion-process approximation considered in the next section.

(3) Borovkov (1984) (describing mainly the weak convergence of queueing processes) contains a wealth of material on asymptotic methods published by probabilists from the former Soviet Union and former East Germany.

8.2 Brownian Motion Process

8.2.1 Introduction

The transient-state distribution of the queue length of even the simple $M/M/1$ queue is difficult to handle; that for the $M/G/1$ queue is not known. In view of such difficulties, some methods of approximation for more general queueing models are considered. One such method that originated with the works of Iglehart (1965), Gaver (1968), and Newell (1971) involves the use of the diffusion process to approximate queue-length distribution under heavy traffic conditions. The idea is to approximate the discrete and random arrivals by a nonrandom continuum and to do the same for the departures. The analogy is fluid flow—fluid flowing into and out of a reservoir. A moment's reflection will convince one about the appropriateness of this approach. During rush hour people coming out of a subway or an electric train from a busy station resemble a continuous flow as opposed to a discrete, random flow in a very lean hour.

There are two ways to view the asymptotic behavior of a queueing process (discrete state). One is to obtain a strong-law-of-large-numbers type of limit, and the other is to obtain a central-limit-theorem type of limit. In the former case, the limit is a deterministic (nonrandom) function of time, while in the latter case, the limit is a stochastic process (a random function of time), specifically a Brownian motion (diffusion) process.

The first type of limit is generally referred to as a FLLN (functional-law-of-large-numbers) limit, and the second as a FCLT (functional Central Limit Theorem). The corresponding models that describe the asymptotic behavior of the queueing process are called, respectively, "fluid model" and "Brownian or diffusion model"; but since both the models have continuous-state space, they are both called a "fluid type" of model. We shall confine ourselves here to the second type of model, leading to diffusion approximation.

We shall now briefly discuss a diffusion process. For details the reader may refer to a standard book on stochastic processes; see also Newell (1971), Kleinrock (1976, **vol. II**), Heyman and Sobel (1982), Gelenbe and Mitrani (1980), and Harrison (1985).

Definition A stochastic process $\{X(t), t \geq 0\}$ satisfying the following properties is called a *Brownian motion process* with drift m and variance parameter D^2.

(i) $X(t)$ has independent increments; that is, for every pair of disjoint time intervals, say, (s, t) and (u, v), $s < t \leq u < v$, the increments $\{X(t) - X(s)\}$ and $\{X(v) - X(u)\}$ are independent random variables. In fact, the definition extends to any $k(\geq 2)$ independent increments (not just $k = 2$ as indicated here).

(ii) Every increment $\{X(t) - X(s)\}$ is normally distributed with mean $m(t-s)$ and variance $D^2(t-s)$.

Property (i) implies that a Brownian motion process is a Markov process. In fact, the property of independent increments is more restrictive than the Markov property, and Property (ii) implies that it is Gaussian. Thus,

$$\begin{aligned} Pr\{X(t) \le x \mid X(s) = x_0\} &= Pr\{X(t) - X(s) \le x - x_0\} \\ &= \Phi(\alpha), \quad \text{where} \end{aligned} \tag{8.2.1}$$

$$\begin{aligned} \alpha &= \frac{\{x - x_0 - m(t-s)\}}{D\sqrt{(t-s)}} \quad \text{and} \\ \Phi(t) &= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \end{aligned} \tag{8.2.2}$$

is the DF of the standard normal variate. The quantities m and D^2 may also be interpreted as follows:

$$m = \lim_{\Delta t \to 0} \frac{E\{X(t+\Delta t) - X(t)\}}{\Delta t}, \tag{8.2.3}$$

that is,

$$\begin{aligned} m\Delta t &= E\{X(t+\Delta t) - X(t)\} + o(\Delta t) \quad \text{and} \\ D^2 &= \lim_{\Delta t \to 0} \frac{E\{X(t+\Delta t) - X(t)\}^2}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\text{var}\{X(t+\Delta t) - X(t)\}}{\Delta t} \\ D^2 \Delta t &= \text{var}\{X(t+\Delta t) - X(t)\} + o(\Delta t). \end{aligned} \tag{8.2.4}$$

The quantities m and D^2 are called the infinitesimal mean and infinitesimal variance, respectively, of the process. We assume here that m and D^2 are constants, independent of t or of x, where $X(t) = x$. By considering m and D^2 as functions of t or of x or of both t and x, we get more general processes, which we shall consider in Section 8.3.6.

Theorem 8.2. *The distribution of* $\{X(t) \mid X(0) = x_0\}$ *for large* t *is exponential with parameter* $-2m/D^2$.

Proof: Let

$$F\{(t, x; x_0)\} = Pr\{X(t) \le x \mid X(0) = x_0\}. \tag{8.2.5}$$

Then it can be shown that F satisfies the diffusion equation (or forward Kolmogorov equation)

$$\frac{\partial}{\partial t}F(t,x;x_0) = -m\frac{\partial}{\partial x}F(t,x;x_0) + \frac{D^2}{2}\frac{\partial^2}{\partial x^2}F(t,x;x_0). \tag{8.2.6}$$

The initial condition $X(0) = x$ gives

$$F(0,x;x_0) = \begin{cases} 0, & x < x_0 \\ 1, & x \geq x_0. \end{cases}$$

With a reflecting barrier placed on the x-axis, the boundary condition is

$$F(t,0;x_0) = 0, \quad x_0 > 0, \quad t > 0. \tag{8.2.7}$$

This is called the "reflecting boundary condition." The solution of the diffusion equation (8.2.6) subject to the preceding conditions is given by

$$F(t,x;x_0) = \Phi\left(\frac{x - x_0 - mt}{D\sqrt{t}}\right) - e^{2xm/D^2}\Phi\left(\frac{-x - x_0 - mt}{D\sqrt{t}}\right). \tag{8.2.8}$$

The preceding result, which holds for both $m > 0$ and $m < 0$, gives a time-dependent solution. We are also interested in a steady-state solution.

Assume that $\lim_{t\to\infty} F(t,x;x_0) = F(x)$; then (8.2.6) reduces to

$$0 = -m\frac{\partial}{\partial x}F(x) + \frac{D^2}{2}F(x). \tag{8.2.9}$$

When $m < 0$

$$\lim_{t\to\infty}\Phi\left(\frac{x - x_0 - mt}{D\sqrt{t}}\right) = \lim_{t\to\infty}\Phi\left(\frac{-x - x_0 - mt}{D\sqrt{t}}\right) = 1,$$

so that from (8.2.8), we have

$$F(x) = 1 - e^{2xm/D^2} = 1 - \exp\left\{(-2x)\frac{(-m)}{D^2}\right\}. \tag{8.2.10}$$

Thus, the distribution of $\{X(t) \mid X(0) = x_0\}$ for large t is exponential with mean

$$\lim_{t\to\infty} E\{X(t) \mid X(0) = x_0\} = \frac{D^2}{-2m}.$$

■

Remarks:

(1) It is evident that $F(t, x; x_0) = \Phi[(x - x_0 - mt)/D\sqrt{t}]$ is a solution of the diffusion equation (8.2.6). But this solution does not satisfy the boundary condition (8.2.7); the solution of (8.2.6) that satisfies the boundary condition (8.2.7) is given by (8.2.8). Though it is not difficult to solve the diffusion equation (8.2.6) by itself, imposition of specific boundary conditions makes its solution difficult. The boundary conditions imposed are to be meaningful. The reflecting boundary condition contained in (8.2.7) is meaningful in the context of the problem studied here.

For a derivation of (8.2.8), readers are referred to Newell (1971) and Kleinrock (1976, **vol. II**).

The diffusion equation (8.2.6) can also be put in terms of the PDF of $\{X(t) \mid X(0) = x_0\}$.

(2) We shall proceed to apply a diffusion process approach in the study of the $G/G/1$ system. The justification of such an approach is provided by an important limit theorem established by Iglehart and Whitt (1970). The limit theorem shows that the queue-length and waiting-time processes of a $G/G/c$ system can be approximated by Brownian motion processes. The importance of the approach increases because of the fact that an exact solution of even the $M/G/1$ queue is not known and, wherever known, exact solutions of still simpler queues are difficult to handle analytically.

8.2.2 Asymptotic queue-length distribution

Let $A(t)$, $D(t)$, and $N(t)$ denote, respectively, the number of arrivals, number of departures, and number in the system at time t of a system. Here $\{A(t), t \geq 0\}, \{D(t), t \geq 0\}$, and $\{N(t), t \geq 0\}$ are stochastic processes. We assume that the system is under heavy traffic and that

$$N(t) = N(0) + A(t) - D(t).$$

The assumption that $N(t)$ does not become zero is basic in this approach. The departure process $\{D(t), t \geq 0\}$, which is otherwise dependent upon the arrival process $\{A(t), t \geq 0\}$, then becomes approximately independent of the arrival process. The number of departures increases by unity each time a service is completed, and the interdeparture time will have the same distribution as service times when the system remains continually busy. Let the IID random variables t_i, $i = 1, 2, \ldots$, denote the interarrival times and let

$$T_n = t_1 + \cdots + t_n.$$

The nth arriving customer arrives at the epoch T_n. We have the important equivalence relation

$$Pr\{A(t) \geq n\} = Pr\{T_n \leq t\}. \tag{8.2.11}$$

The preceding relation enables us to find the distribution of $A(t)$ from that of T_n. Since t_is are IID random variables, the Central Limit Theorem can be applied to find the asymptotic distribution of T_n. We have

$$E\{T_n\} \simeq \frac{n}{\lambda}$$
$$\text{var}\{T_n\} \simeq n\sigma_u^2$$

where $1/\lambda$ and σ_u are the mean and SD of the interarrival times, respectively.

From the Central Limit Theorem, we have

$$Pr\left\{\frac{T_n - \frac{n}{\lambda}}{\sigma_u\sqrt{n}} \leq x\right\} = \Phi(x). \tag{8.2.12}$$

To find the RHS of (8.2.8), we have to relate n with t. Define

$$t = x\sigma_u\sqrt{n} + \frac{n}{\lambda}.$$

For large n, the dominant term being $t \simeq n/\lambda$, we can express n in terms of t as follows:

$$n \simeq \lambda t - x\lambda\sigma_u\sqrt{t\lambda}.$$

From (8.2.12) we have

$$Pr\{T_n \leq t\} = \Phi(x)$$

so that

$$\begin{aligned}
Pr\{A(t) \geq n\} &= \Phi(x) \quad \text{or} \\
Pr\{A(t) \geq \lambda t - x\lambda\sigma_u\sqrt{t\lambda}\} &= \Phi(x) \quad \text{or} \\
Pr\left\{\frac{A(t) - \lambda t}{\lambda\sigma_u\sqrt{t\lambda}} \geq -x\right\} &= \Phi(x) \quad \text{or} \\
Pr\left\{\frac{A(t) - \lambda t}{\lambda\sigma_u\sqrt{t\lambda}} \leq x\right\} &= 1 - \Phi(-x) = \Phi(x).
\end{aligned} \tag{8.2.13}$$

Thus, the asymptotic distribution of $A(t)$ is Gaussian, with

$$\begin{aligned}
E\{A(t)\} &\simeq \lambda t \\
\text{var}\{A(t)\} &\simeq \lambda^3\sigma_u^2 t.
\end{aligned} \tag{8.2.14}$$

Denoting the mean and SD of service-time distribution by $1/\mu$ and σ_v, we can show that $D(t)$ is also asymptotically normal with

$$\begin{aligned}
E\{D(t)\} &\simeq \mu t \quad \text{and} \\
\text{var}\{D(t)\} &\simeq \mu^3\sigma_u^2 t.
\end{aligned} \tag{8.2.15}$$

That is,

$$D(t) \sim N(\mu t, \mu^3 \sigma_v^2 t).$$

The result can be put as follows.

Theorem 8.3. *For large t and for moderate to heavy-loaded queueing systems it can be said that*

$$N_1(t) = N(t) - N(0) = A(t) - D(t)$$

is a Gaussian process with

$$E\{N_1(t)\} \simeq \lambda t - \mu t = \mu(\rho - 1)t \quad \textit{and} \tag{8.2.16}$$

$$\text{var}\{N_1(t)\} \simeq (\lambda^3 \sigma_u^2 + \mu^3 \sigma_v^2)t. \tag{8.2.17}$$

Remarks:

(1) It is suggested that the process $\{N(t), t \geq 0\}$, where

$$N(t) = N(0) + A(t) - D(t)$$

can be approximated by a diffusion process having infinitesimal mean m and variance D^2 given by

$$m = \lim_{\Delta t \to 0} \frac{E\{N(t + \Delta t) - N(t)\}}{\Delta t} = \lambda - \mu \tag{8.2.18}$$

$$D^2 = \lim_{\Delta t \to 0} \frac{\text{var}\{N(t + \Delta t) - N(t)\}}{\Delta t} = \lambda^3 \sigma_u^2 + \mu^3 \sigma_v^2. \tag{8.2.19}$$

Equation (8.2.19) follows from the fact that $\{N(t), t \geq 0\}$ has independent increments and that $\text{cov}\{N(t), N(s)\} = \text{var}\{N[\min(t, s)]\}$ holds for such a process.

(2) The expressions (8.2.16) and (8.2.18) give a reasonable approximation of the mean when $\rho > 1$; however for $\rho < 1$, (8.2.16) becomes negative; this defect is taken care of by considering a reflecting barrier for $N(t)$ at the origin.

(3) That (8.2.14) and (8.2.15) hold for large t also follows from a result of renewal theory (Cox 1962). For a renewal process $\{R(t), t \geq 0\}$ where the interrenewal times have mean $1/v$ and variance σ^2, we have, for large t,

$$\begin{aligned} E\{R(t)\} &\simeq vt \quad \text{and} \\ \text{var}\{R(t)\} &\simeq \sigma^2 v^3 t. \end{aligned} \tag{8.2.20}$$

Here $\{A(t), t \geq 0\}$ is a renewal process. For ρ close to 1, $\{D(t), t \geq 0\}$ can also be approximated as a renewal process.

(4) The results, though based on a Central Limit Theorem approach, are renewal theoretic results.

8.2.3 Diffusion approximation for a $G/G/1$ queue

Let

$$F(t, x; x_0) = Pr\{N(t) \leq x \mid N(0) = x_0\}.$$

The Brownian motion process $\{N(t), t \geq 0\}$ has the drift m and variance D^2. The assumption is that $N(t) > 0$. The lower boundary at $x = 0$ for $\{N(t), t \geq 0\}$ would act as a reflecting barrier.

Theorem 8.4. *Diffusion approximation of the steady-state queue-length distribution $\{\hat{p}_n\}$ in a $G/G/1$ queue is given by*

$$\begin{aligned} \hat{p}_n &= \rho(1-\hat{\rho})(\hat{\rho})^n, \quad n \geq 1 \\ \hat{p}_0 &= 1-\rho, \end{aligned} \tag{8.2.21}$$

where

$$\begin{aligned} \hat{\rho} = e^{-\gamma} &= e^{2m/D^2} \\ &= \exp\left\{-\frac{2(1-\rho)}{\mu^2(\rho^3\sigma_u^2 + \sigma_v^2)}\right\}. \end{aligned} \tag{8.2.22}$$

Proof: We find from (8.2.10) that the steady-state distribution

$$\begin{aligned} F(x) &= \lim_{t\to\infty} F(t, x; x_0) \\ &= 1 - \exp\left\{-\frac{2x(-m)}{D^2}\right\}, \quad x \geq 0, \end{aligned}$$

so that the distribution of the number in the system in steady state is exponential with

$$\begin{aligned} \text{mean} &= \lim_{t\to\infty} E\{N(t) \mid N(0) = x_0\} \\ &= \frac{D^2}{(-2m)} \\ &= \frac{1}{\gamma} \quad \text{(say)}. \end{aligned}$$

Putting the values of m and D^2 as found in (8.2.18) and (8.2.19), we get that the distribution of $N(t)$ is exponential with mean

$$\frac{1}{\gamma} = \frac{D^2}{-2m} = \frac{\mu^2(\rho^3\sigma_u^2 + \sigma_v^2)}{2(1-\rho)}. \tag{8.2.23}$$

Kobayashi (1974) suggests that we may discretize the distribution. Thus, the steady-state distribution of a number N in the system is given by

$$\begin{aligned}\hat{p}_n &\equiv Pr\{N = n\} = \int_n^{n+1} dF(x) \\ &= F(n+1) - F(n) \\ &= e^{-n\gamma}(1 - e^{-\gamma}) \\ &= (1 - \hat{\rho})(\hat{\rho})^n, \quad n = 0, 1, 2, \ldots, \end{aligned} \tag{8.2.24a}$$

where

$$\hat{\rho} = e^{-\gamma} \quad \text{(as in (8.2.22))}.$$

This is expected, since discretization of exponential distribution leads to geometric distribution.

Kobayashi further suggests that one needs to choose a boundary condition from the simple reflecting barrier. We get $1 - \hat{p}_0 = \hat{\rho}$, whereas the exact value of server utilization is ρ. He therefore recommends that the probability that the system is empty be taken as $(1 - \rho)$ and the result (8.2.24a) be modified as follows:

$$\begin{aligned}\hat{p}_0 &= 1 - \rho \\ \hat{p}_n &= \rho(1 - \hat{\rho})(\hat{\rho})^{n-1}, \quad n = 1, 2, \ldots. \end{aligned} \tag{8.2.24b}$$

We thus get the diffusion approximation of the steady-state distribution of the number in the system for moderately high to heavy traffic. ■

8.2.3.1 Particular cases

M/G/1 System

Here $\sigma_u^2 = 1/\lambda^2$ so that

$$\begin{aligned}\hat{L} = \lim_{t\to\infty} E\{N(t)\} &= \frac{D^2}{-2m} \\ &= \frac{\left(\frac{\lambda^2\sigma_v^2 + \rho^3}{\rho^2}\right)}{2(1-\rho)}. \end{aligned} \tag{8.2.25}$$

From the Pollaczek-Khinchin formula, we get, using $L = \lambda W$, that

$$\begin{aligned}L = \lim_{t\to\infty} E\{N(t)\} &= \rho + \frac{\lambda^2\sigma_v^2 + \rho^2}{2(1-\rho)} \\ &= \frac{\lambda^2\sigma_v^2 + (2\rho - \rho^2)}{2(1-\rho)}. \end{aligned} \tag{8.2.26}$$

The RHS of (8.2.25) and (8.2.26) are both close when ρ is close to 1. The error in approximating the mean number in the system in equilibrium is small when ρ is close to 1.

M/M/1 System

Here $\sigma_u^2 = 1/\lambda^2$ and $\sigma_v^2 = 1/\mu^2$. so that

$$\hat{\rho} = \exp\left\{\frac{-2(1-\rho)}{(1+\rho)}\right\}.$$

This value of $\hat{\rho}$ is close to ρ in the neighborhood of $\rho = 1$. When $\rho \to 1$ $\hat{\rho} \to 1$ so that (8.2.24) becomes

$$\hat{\rho}_n \simeq (1-\hat{\rho})(\hat{\rho})^n, \quad n = 0, 1, 2, \ldots;$$

and the approximation of $\hat{\rho}_n$ is good when ρ is close to 1 (and slightly less than 1).

8.2.4 Virtual delay for the G/G/1 system

Let $X(t)$ denote the total workload by time t; that is, $X(t)$ is the total time required by the server to complete serving all units that arrived in the interval $(0, t]$. Suppose that $X(0) = 0$. We have

$$X(t) = v_1 + v_2 \cdots + v_{A(t)}.$$

Let $W(t)$ denote the remaining workload (or work backlog)—that is, the time required by the server to complete serving all units present at the epoch t. $W(t)$ is called the virtual waiting time—that is, the time that an imaginary customer would have to wait in the queue were he or she to arrive at the epoch t.

$W(t)$ is the sum of the residual service time at epoch t (of the unit being served) plus the service time of units waiting at epoch t.

Suppose that $W(0) = 0$. If the server has been busy throughout the interval $(0, t]$ and if he or she works continuously at a unit rate in that interval, then we have

$$\begin{aligned} W(t) &= W(0) + X(t) - t \\ &= X(t) - t \end{aligned} \tag{8.2.27}$$

when $W(0) = 0$. When $W(t) \geq \delta t$,

$$W(t+\delta t) - W(t) = X(t+\delta t) - X(t) - \delta t. \tag{8.2.28}$$

However, the simple relation given in (8.2.27) does not hold good in the general case. Nevertheless, it can be seen that (8.2.28) holds good in the general case whenever $W(t) \geq \delta t$. A look at Fig. 8.1 will convince the reader about the validity of (8.2.28). We have from the figure, $W(t) > \delta t$ and

$$W(t+\delta t) - W(t) = EF.$$

Again,

$$X(t+\delta t) - X(t) - \delta t = DB - CF = FH = EF$$

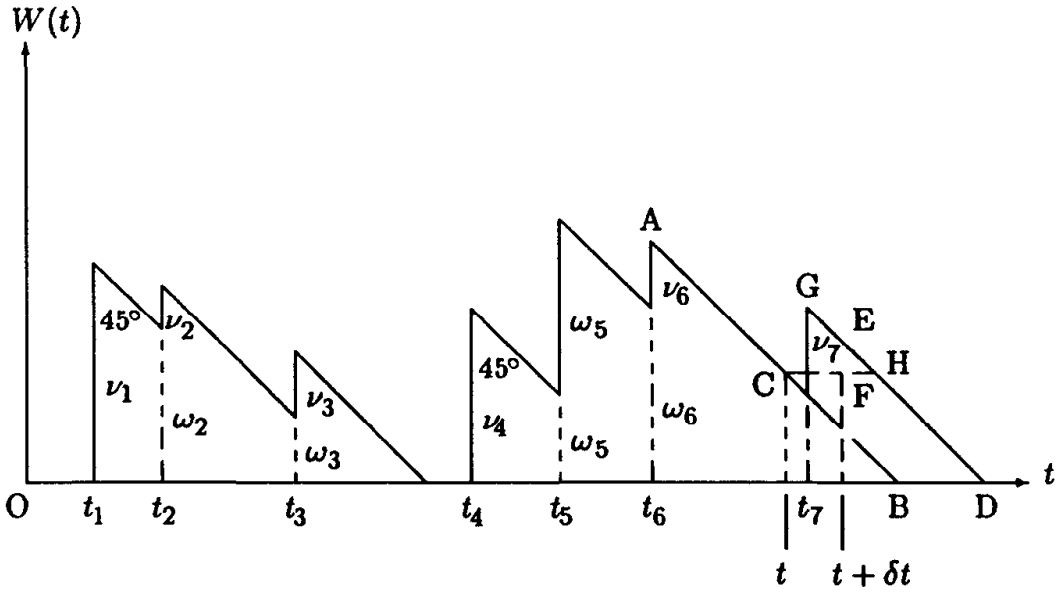


Figure 8.1 Graph of the virtual waiting time $W(t)$.

so that (8.2.28) holds. Taking expectation, we get

$$E\{X(t)\} = E(v_i)E\{A(t)\} \simeq \frac{1}{\mu}\lambda t = \rho t. \tag{8.2.29a}$$

Using the relation

$$\text{var}\{X(t)\} = \text{var}(v_i)E\{A(t)\} + \text{var}\{A(t)\}\{E(v_i)\}^2,$$

we get

$$\text{var}\{X(t)\} \simeq \sigma_v^2\lambda t + \lambda^3\sigma_u^2 t\left(\frac{1}{\mu}\right)^2 = \lambda\left(\sigma_v^2 + \rho^2\sigma_u^2\right)t. \tag{8.2.29b}$$

Assume that (8.2.29a) and (8.2.29b) hold for all t. We have

$$\lim_{\delta t\to 0}\frac{E\{X(t+\delta t) - X(t)\}}{\delta t} = \rho \quad \text{and}$$
$$\lim_{\delta t\to 0}\frac{\text{var}\{X(t+\delta t) - X(t)\}}{\delta t} = \lambda\left(\sigma_v^2 + \rho^2\sigma_u^2\right). \tag{8.2.30}$$

Using (8.2.28), we get

$$\lim_{\delta t\to 0}\frac{E\{W(t+\delta t) - W(t)\}}{\delta t} = \rho - 1 \quad \text{and} \tag{8.2.31}$$
$$\lim_{\delta t\to 0}\frac{\text{var}\{W(t+\delta t) - W(t)\}}{\delta t} = \lim_{\delta t\to 0}\frac{\text{var}\{X(t+\delta t) - X(t)\}}{\delta t} = \lambda\left(\sigma_v^2 + \rho^2\sigma_u^2\right). \tag{8.2.32}$$

The mean and variance of increments in $W(.)$ are proportional to δt for small δt. This is a property of the Brownian motion process. We can approximate the virtual-waiting-time process $\{W(t), t \geq 0\}$ by a Brownian motion process having a reflecting barrier at the x-axis. The parameters of this process are

$$\begin{aligned} m &= \rho - 1 \quad \text{and} \\ D^2 &= \lambda(\sigma_v^2 + \rho^2 \sigma_u^2). \end{aligned} \tag{8.2.33}$$

Thus,

$$F(t, x; x_0) = Pr\{W(t) \leq x \mid W(0) = x_0\} \tag{8.2.34}$$

is given by (8.2.8) and in the limit, as $t \to \infty$

$$F(x) \simeq 1 - e^{-2x(-m)/D^2}. \tag{8.2.35}$$

Thus, we have the following result.

Theorem 8.5. *The diffusion approximation of the limiting distribution of the virtual waiting time $W(t)$ in a $G/G/1$ queue is exponential with mean*

$$\begin{aligned} \lim_{t\to\infty} E\{W(t)\} &= \lim_{t\to\infty} E\{W(t) \mid W(0) = x_0\} \\ &= \frac{D^2}{-2m} \\ &= \frac{\lambda(\sigma_v^2 + \rho^2\sigma_u^2)}{2(1-\rho)}. \end{aligned} \tag{8.2.36}$$

8.2.4.1 Particular case: $M/G/1$ system

When arrivals occur in accordance with a Poisson process, $\sigma_u^2 = 1/\lambda^2$. From (8.2.36), we then get

$$\lim_{t\to\infty} E\{W(t)\} = \frac{\lambda(\sigma_v^2 + \frac{1}{\mu^2})}{2(1-\rho)}, \tag{8.2.37}$$

which is the Pollaczek-Khinchin mean-value formula (for mean waiting time in the queue).

Notes:

(1) The steady-state distribution of the virtual waiting time is identical with the steady-state distribution of the actual waiting time of an arriving customer if and only if arrivals occur in accordance with a Poisson process. When the input is other than Poisson, the two steady-state distributions are different.

(2) For Poisson input, $X(t)$ is a compound Poisson process, with DF

$$Pr\{X(t) \leq x\} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(-\lambda t)^n}{n!} B^{n*}(x),$$

where $B^{n*}(x)$ is the n-fold convolution of service time.

8.2.5 Approach through an absorbing barrier with instantaneous return

In obtaining the results in Section 8.2.3, it was assumed that the lower boundary acts as a reflecting barrier for the process $\{N(t), t \geq 0\}$, where $N(t)$ is the number in the system at time t. Gelenbe (1979) has considered an alternative approach through the *instantaneous returns process*. The stochastic process $\{N(t), t \geq 0\}$ is considered to represent the position of a particle moving on a closed interval $(0, \infty]$ of the real line. When the particle reaches the lower boundary $x = 0$ of the interval, it remains there for a period of time (denoted by the RV ξ) at the end of which the particle jumps instantaneously back to the open interval $[0, \infty]$ to a random point. The origin is an absorbing barrier for a length ξ. The RV ξ corresponds to the idle period. Let $E(\xi) = 1/\lambda'$. Note that for a Poisson input $\lambda' = \lambda$. Gelenbe shows that the PDF of the steady-state distribution is given by

$$f(x) = \begin{cases} R(e^{\gamma} - 1)e^{-\gamma x}, & x \geq 1 \\ R(1 - e^{-\gamma x}), & 0 \leq x \leq 1, \end{cases} \tag{8.2.38}$$

where γ is as given in (8.2.23) and

$$R = \frac{\lambda'}{\lambda' + \mu - \lambda}, \tag{8.2.39}$$

the condition for the existence of a steady-state solution being $\rho = \lambda/\mu < 1$.

When the input is Poisson, $R = \rho = \lambda/\mu$. Then for an $M/G/1$ queue the expected number in the steady state is approximately given by

$$\begin{aligned} E(N) &= \int_0^{\infty} x f(x) dx \\ &= \rho\left(\frac{1}{2} + \frac{1}{\gamma}\right) \\ &= \rho\left(\frac{1}{2} + \frac{\rho + \mu^2\sigma_v^2}{2(1-\rho)}\right). \end{aligned} \tag{8.2.40}$$

This is close to $E(N)$ for an $M/G/1$ given by the Pollaczek-Khinchin formula.

8.2.6 Diffusion approximation for a $G/G/c$ queue: state-dependent diffusion equation

The idea of diffusion approximation for a $G/G/1$ model can be extended in principle to a multiserver $G/G/c$ model. The input process $\{A(t), t \geq 0\}$ of a multiserver system is not in any way affected by the number of servers. Thus, (8.2.14) will be valid for a c-server system—that is,

$$E\{A(t)\} \simeq \lambda t \quad \text{and}$$
$$\text{var}\{A(t)\} = \lambda^3 \sigma_u^2 t.$$

However, Eq. (8.2.15) for the departure process $\{D(t), t \geq 0\}$ have to be suitably modified. When $n < c$, the departure rate will be $n\mu$, and we shall have

$$\begin{aligned} E\{D(t)\} &\simeq n\mu t \quad \text{and} \\ \text{var}\{D(t)\} &\simeq n\mu^3 \sigma_v^2 t; \end{aligned} \tag{8.2.41}$$

and for $n \geq c$, the departure rate will be $c\mu$ so that

$$\begin{aligned} E\{D(t)\} &\simeq c\mu t \quad \text{and} \\ \text{var}\{D(t)\} &\simeq c\mu^3 \sigma_v^2 t. \end{aligned} \tag{8.2.42}$$

We can apply similar arguments as in the case of a single-server model and use diffusion approximation. The infinitesimal mean and variance will be state-dependent, and, in place of those of (8.2.18) and (8.2.19), we shall have state-dependent m and D^2 as functions of x, where $X(t) = x$. Thus,

$$m \equiv b(x) = \lambda - [\min\{x, c\}]\mu \tag{8.2.43}$$
$$D^2 \equiv a(x) = \lambda^3 \sigma_u^2 + [\min\{x, c\}]\mu^3 \sigma_v^2. \tag{8.2.44}$$

When $c = 1$, these reduce to (8.2.18) and (8.2.19), respectively.

8.2.6.1 State-dependent diffusion equation

Define the PDF of conditional $X(t)$ by

$$f \equiv f(t, x; x_0) = Pr\{x \leq X(t) < x + dx \mid X(0) = x_0\}. \tag{8.2.45}$$

Then it can be shown that $f(t, x, x_0) \equiv f$ satisfies the diffusion equation (forward Kolmogorov equation)

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial x}\{b(x) f\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{a(x) f\}, \tag{8.2.46}$$

where the diffusion parameters are given by

$$b(x) = \lim_{\Delta t \to 0} \frac{E\{X(t+\Delta t) - X(t) \mid X(t) = x\}}{\Delta t} \quad \text{and} \tag{8.2.47}$$

$$a(x) = \lim_{\Delta t \to 0} \frac{\text{var}\{X(t+\Delta t) - X(t) \mid X(t) = x\}}{\Delta t}. \tag{8.2.48}$$

Note that we have put the equation in terms of the PDF instead of the DF. We now make appropriate choices of the diffusion parameters and the boundary conditions.

As discussed already for a $G/G/c$ system, $b(x)$ and $a(x)$ can be taken as given in (8.2.43) and (8.2.44).

8.2.6.2 Steady-state solution

Let

$$\lim_{t \to \infty} f(t, x; x_0) = f(x).$$

Then Eq. (8.2.46) reduces to

$$\frac{d}{dx}\{b(x) f(x)\} = \frac{1}{2}\frac{d^2}{dx^2}\{a(x) f(x)\}. \tag{8.2.49}$$

By imposing appropriate boundary conditions, Eq. (8.2.49) can be solved to yield $f(x)$. Heyman and Sobel (1982) give a method of solving (8.2.49) with the boundary condition

$$2b(0) f(0) = \frac{d}{dx}\{a(x) f(x)\}|_{x=0}. \tag{8.2.50}$$

They also make a comparison between the exact values and the diffusion-approximation values (as obtained by the method) for the mean and variance of queue length for an $M/M/c$ queue for certain values of c. The agreement is quite satisfactory.

8.2.7 Diffusion approximation for an $M/G/c$ model

Analytical solutions of an $M/G/c$ queue for distributions of queue length, waiting time, and busy period are not known. Thus, approximate methods assume importance. Various methods, including the diffusion-process approach, have been put forward. Because of certain special properties of the Poisson input process, one can expect to get sharper results than for a $G/G/c$ queue through diffusion approximation.

Halachmi and Franta (1978) and Sunaga *et al.* (1978) investigated the $M/G/c$ queue on the basis of the heavy-traffic-limit theorem. Halachmi and Franta

took as diffusion parameters the functions given in (8.2.43) and (8.2.44) and reflecting barrier as boundary condition at the origin.

Kimura (1983) proposes as diffusion parameters

$$\begin{aligned} b(x) &= \lambda - \{\min([x], c)\}\mu \quad \text{and} \\ a(x) &= \lambda + \{\min([x], c)\}\mu^3\sigma_v^2, \end{aligned} \tag{8.2.51}$$

where $[x]$ is the smallest integer not smaller than x. The infinitesimal mean and variance (given in (8.2.51)) of the diffusion process $\{X(t), t \geq 0\}$ are piecewise continuous.

As a boundary condition, Kimura treats the origin as an "elementary return boundary," which implies (in a rough sense) that when the process $\{X(t), t \geq 0\}$ reaches the state 0, it stays there for a random length of time ξ whose distribution is exponential and then returns to the interval $[0, \infty]$. It may be seen that for a Poisson input process, the intervals during which the system is empty (idle periods) are exponentially distributed and that from $X(.) = 0$, the process jumps to $X(.) = 1$ with the arrival of a fresh customer. This boundary has also been used by Gelenbe and others. Using some other properties of the Poisson input queue (such as PASTA), Kimura obtains approximate formulas for the steady-state distributions of the queue length, the waiting time, and the busy period. It is shown that the formula of the mean waiting time obtained by Kimura through diffusion approximation agrees when $\rho \to 1$ with that obtained by Köllerström (1974) through heavy-traffic approximation.

Kimura and Ohsone (1984) extend the results to the system $M^{[x]}/G/c$ with the group arrivals. Yao (1985) suggests some refinements to the diffusion approximation of the $M/G/c$ system by Kimura (1983). Yao retains the piecewise continuity formulation of the diffusion parameters but suggests a modification of the boundary condition: replacement of the elementary return boundary by a reflecting barrier. He uses an alternative approach for the solution of the diffusion equation and incorporates some known results into the model to obtain his solution. Some numerical comparisons are also given.

8.2.8 Concluding remarks

Though the approach of approximating a discrete-state (or jump) process by a diffusion process with continuous path is not new, application of the principle to queueing theory is of rather recent origin. The method of approximation has gained importance because of the mathematical intractability of most of the queueing models on the one hand and the satisfactory nature of the approximating solution on the other hand.

We have restricted our discussion to a steady-state type of solution of queueing models. Transient solutions, which are much more difficult to obtain, have also been attempted through diffusion approximation. (See Newell (1971) and Kobayashi (1974) for an account.)

Another direction in which the method of diffusion approximation has been applied and found suitable is the analysis of open and close networks of queues. Here one has to deal with a multidimensional diffusion equation. (See Kobayashi (1974), **part II**, and Reiser and Kobayashi (1974).)

The diffusion-process approximation has been found useful in application of queueing theory in several areas (in modeling computer and communication systems; see Reiser (1982) and Chung and Williams (1990).

8.3 Queueing Systems with Vacations

8.3.1 Introduction

It may happen in several situations that the server is unavailable to the customers (primary customers) over occasional periods of time. The server may then be doing other work, such as maintenance work or servicing secondary customers. The periods for which the server is unavailable are said to be server-vacation periods. Systems with server vacations can be used as models of many production, communication, and computer systems. Two examples are given next. Doshi (1986) gives a large number of examples. See also Takagi (1991).

(i) *Production systems.* Machines producing certain items may need periodic checking and maintenance. The periods of random lengths of preventive maintenance may be considered as periods of server vacation when the server is unavailable.

(ii) *Computer and communication systems.* A server in such a system, besides being engaged in primary functions (such as receiving, processing, and transmitting data), has to undertake secondary works such as preventive maintenance or has to scan for new work for occasional periods of time. There are several other situations when the server is unavailable to primary customers for occasional (and random) periods of time when there is a lull in data traffic.

Vacation models have of late received much attention for their interesting theoretical properties as well as for applicability in more complicated queueing (or polling) models. These have applications in telecommunications and computer network design. See, for example, Takagi (1991), Bertsekas and Gallager (1992), and so on.

There are different service disciplines that could be considered in a vacation model, such as exhaustive service, gated service, different kinds of limited service, and decrementing service.

Vacations with exhaustive service

We make the following assumptions. Each time a busy period ends and the system becomes empty, the server starts a vacation of random length of

time. When the server returns from vacation and finds one or more customers waiting, he goes on serving until the system becomes empty (called *exhaustive service* discipline). If on return from a vacation (at the end of a busy period), the server finds no customer waiting, he waits for the arrival of a customer. This is called a *single-vacation system* and is denoted by V_s. On the other hand, if he finds no customer waiting, he goes on taking vacations until, on return from a vacation, he finds at least one customer waiting. This is called *a multiple-vacation system* and is denoted by V_m. We denote the nth vacation by v_n and assume that $\{v_n, n = 1, 2, \ldots\}$ is a sequence of IID random variables independent of the service time; the sequence $\{v_n\}$ of RVs may be independent of the arrival period or may be dependent on the arrival process. Let F_v be the DF of v_n and $\tilde{f}_v$ be the LST of F_v. Consider a standard $G/G/1$ queue with interarrival time A (with mean $1/\lambda$ and SD σ_u) and service time S (with mean $1/\mu$ and SD σ_v). We shall denote a queue with general interarrival and service times with single vacation by $G/G/1 - V_s$ and the corresponding queue with multiple vacation by $GI/G/1 - V_m$.

Starting with Gaver (1962), vacation queues have been receiving attention from researchers. Doshi (1986) gives a survey (see also Doshi (1990)). Takagi (1991) deals with vacation models with Poisson input in detail.

8.3.2 Stochastic decomposition

It has been observed that the queues with server vacation exhibit an interesting property called the stochastic decomposition property. Under certain conditions on the sequence $\{v_n\}$, both for $G/G/1 - V_s$ and $G/G/1 - V_m$ models, the steady-state waiting time is the sum of two independent random variables. One is the waiting time in the same $G/G/1$ queue without vacation, and the other is a random variable related to $\{v_n\}$.

For queues with Poisson input, such a decomposition property holds even for the steady-state queue-length distribution.

The decomposition property for Poisson input systems has been observed, for example, by Gaver (1962), Cooper (1970), Furhrmann (1984), and Furhmann and Cooper (1985), while that for $G/G/1$ vacation models has been discussed by Gelenbe and Iasnogordski (1980), and Doshi (1985). See also Takine and Hasegawa (1992).

8.3.3 Poisson input queue with vacations: [exhaustive-service] queue-length distribution

We shall now discuss the $M/G/1$ queue with single and multiple vacations under exhaustive-service discipline. Assume that the vacation sequence v_n is stationary and that the system is in steady state.

Let N^* be the number of customers present at the start of a busy period following a vacation or vacation period. Clearly, $N^* \geq 1$; N^* can be deterministic or a random variable.

First, consider that N^* is an RV having PGF

$$R(z) = \sum_{n=1}^{\infty} Pr\{N^* = n\} z^n, \quad |z| < 1.$$

Let $P(z)$ be the PGF of the number in the system at a departure epoch of a usual $M/G/1$ queue *without* vacation. Note that the distribution of the number in the system at the random epoch, at an arrival epoch, or at a departure epoch are one and the same for a Poisson input queue (PASTA). $P(z)$ is given by the Pollaczek-Khinchin formula. Let $Q(z)$ be the PGF of the number in the system at a departure epoch of an $M/G/1$ queue with vacations.

Let $V(z)$ be the PGF of the number in the system at a random point in time when the server is on vacation. For a Poisson input queue, the basic decomposition result is

$$Q(z) = P(z)V(z). \tag{8.3.1}$$

(Fuhrmann and Cooper, 1985; Ali and Neuts, 1984)

The basic decomposition result shows that the number of customers at a departure epoch of a Poisson input queue with vacations is the sum of two random variables: (i) the number of customers at a departure epoch at the corresponding Poisson input queue without vacation and (ii) the number of customers at a random point of time given that the server is on vacation.

While variable (i) is vacation-independent, variable (ii) is vacation-related. We now state the important decomposition result (without proof) and consider some special cases. For a proof, refer to Fuhrmann (1984) and Doshi (1986).

Theorem 8.6. *For an $M/G/1$ queue with server vacations,*

$$Q(z) = P(z)\frac{1 - R(z)}{(1 - z)E(N^*)}. \tag{8.3.2}$$

We consider some special cases.

(A) N^* is deterministic

(i) $Pr\{N^* = 1\} = 1$, we get the usual queue with the vacation period corresponding to the idle period of the system. Then $Q(z) = P(z)$.

(ii) N^* is a fixed number, say N—that is, $Pr\{N^* = N\} = 1$. This corresponds to the case when the server is on vacation (or remains busy with other work or secondary customers) until the (primary) queue size builds up to a preassigned fixed number N, known as N-policy, this was considered first by Heyman (1968), who shows that a system with such a policy possesses some optimal properties.

From (8.3.2) we get

$$Q(z) = P(z)\frac{1 - z^N}{(1 - z)N}. \tag{8.3.3}$$

In the preceding two cases under (***A***), the length of the server vacation depends on the arrival process during but not after the vacation. Under (***B***) (given below), the length of server vacation is independent of the arrival process.

(B) N^* is an RV

(a) $M/G/1 - V_m$ system

Let A_v be the number of arrivals during a typical vacation period v. Then the PGF $\alpha(z)$ of A_v is given by

$$\begin{aligned}\alpha(z) &= \sum_{n=0}^{\infty} Pr\{A_v = n\}z^n \\ &= \bar{f}_v[\lambda(1 - z)].\end{aligned} \tag{8.3.4}$$

We have

$$Pr\{A_v = 0\} = \bar{f}_v(\lambda) \tag{8.3.5}$$

so that

$$Pr\{A_v \geq 1\} = 1 - \bar{f}_v(\lambda). \tag{8.3.6}$$

Now the event $N^* = n$ is the event that the number of arrivals during the last vacation period equals n, given that this number is at least 1—that is,

$$\begin{aligned}Pr\{N^* = n\} &= Pr\{A_v = n \mid A_v \geq 1\}, \quad n = 1, 2, \ldots \\ &= \frac{Pr\{A_v = n\}}{1 - \bar{f}_v(\lambda)}.\end{aligned} \tag{8.3.7}$$

Thus,

$$\begin{aligned}R(z) &= \sum_{n=1}^{\infty} Pr\{N^* = n\}z^n \\ &= \frac{\bar{f}_v[\lambda(1 - z)] - \bar{f}_v(\lambda)}{1 - \bar{f}_v(\lambda)}.\end{aligned} \tag{8.3.8}$$

We have

$$E(N^*) = R'(1) = \frac{-\lambda \bar{f}_v'(0)}{1 - \bar{f}_v(\lambda)} = \frac{\lambda E(v)}{1 - \bar{f}_v(\lambda)}.$$

Substituting in (8.3.2), we get

$$Q(z) = P(z)\frac{1 - \tilde{f}_v[\lambda(1-z)]}{\lambda E(v)(1-z)}. \tag{8.3.9}$$

Remark 1. The second factor has an interesting interpretation. Let $Z(t)$ be the forward recurrence time of a vacation (or residual lifetime of the vacation) random variable. Then the limiting distribution Z of $Z(t)$ as $t \to \infty$ is given by

$$F_z(x) = Pr\{Z \le x\} = \frac{\int_0^x [1 - F_v(y)]dy}{E(v)}, \tag{8.3.10}$$

where $F_v(y) = Pr(v \le y)$. (See Eq. (1.7.9) in Ch. 1.) Let b_n be the probability that n arrivals occur during Z and let

$$\beta(z) = \sum_{n=0}^{\infty} b_n z^n$$

be the PGF of the number of arrivals during Z. Then

$$\begin{aligned}
\beta(z) &= \sum_{n=0}^{\infty} z^n \int_0^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} dF_z(t) \\
&= \int_0^{\infty} \left\{ \sum \frac{e^{-\lambda t}(\lambda t z)^n}{n!} \right\} \frac{1 - F_v(t)}{E(v)} dt \\
&= \int_0^{\infty} \frac{e^{-\lambda t(1-z)}\{1 - F_v(t)\}}{E(v)} dt \\
&= \frac{1 - \tilde{f}_v[\lambda(1-z)]}{\lambda E(v)(1-z)},
\end{aligned} \tag{8.3.11}$$

which is equal to the second factor on the RHS of (8.3.9). Thus, while the first factor is the PGF of the number at departure epoch in the standard $M/G/1$ queue without vacation, the second factor is the PGF of the number of arrivals during the limiting forward recurrence time of the vacation period (residual vacation period).

Remark 2. We have $\alpha'(1) = E(A_v) = \lambda E(v)$, so that the second factor on the RHS of (8.3.9) can be written as

$$\frac{1 - \alpha(z)}{(1-z)\alpha'(1)}.$$

Thus, for $M/G/1 - V_m$, (8.3.2) can be put as

$$Q(z) = P(z)\frac{1 - \alpha(z)}{(1-z)\alpha'(1)}, \tag{8.3.12}$$

$\alpha(z)$ being the PGF of the number of arrivals during the vacation.

Note: The factor

$$\frac{1-\alpha(z)}{(1-z)\alpha'(1)}$$

is the PGF of $Pr\{A_v > k\}/E\{A_v\}$ $k = 0, 1, 2, \ldots$.
This is also the PGF of the number of units that arrive during an interval from the commencement of a vacation period to a random point in the vacation period.

Remark 3. In an $M/G/1 - V_m$ queue, the (server) idle period I has mean

$$E(I) = E(v)/[1 - \bar{f}_v(\lambda)].$$

Remark 4. For $M^X/G/1 - V_m$ system, $Q(z)$ will become

$$Q(z) = V(z)\frac{1 - \bar{f}_v\{1 - \lambda A(z)\}}{\lambda E(v)\{1 - A(z)\}},$$

where $A(z) =$ PGF of X, and

$$V(z) = \frac{(1-\rho)(1-z)B^*(\lambda - \lambda A(z))}{B^*(\lambda - \lambda A(z))}.$$

(b) $M/G/1 - V_s$ model
Here there is only one vacation, and there may be no arrivals or one arrival or more than no one arrival during the server-vacation period. If there is no arrival, the server waits for an arrival to occur, and then $N^* = 1$. If there is an arrival during the vacation, then N^* is equal to the number of arrivals during the vacation. Thus,

$$Pr\{N^* = 1\} = Pr\{A_v = 0\} + Pr\{A_v = 1\}$$
$$Pr\{N^* = n\} = Pr\{A_v = n\}, \quad n = 2, 3, \ldots.$$

Thus, using (8.3.4) and (8.3.5), we get

$$\begin{aligned} R(z) &= \sum_{n=1}^{\infty} Pr\{N^* = n\} z^n \\ &= Pr\{A_v = 0\} z + \sum_{n=1}^{\infty} Pr\{A_v = n\} z^n \\ &= z\bar{f}_v(\lambda) + \bar{f}_v[\lambda(1-z)] - \bar{f}_v(\lambda) \\ &= \bar{f}_v[\lambda(1-z)] - (1-z)\bar{f}_v(\lambda) \quad \text{and} \qquad (8.3.13) \\ E(N^*) &= R'(1) = -\lambda \bar{f}_v'(0) + \bar{f}_v(\lambda) \\ &= \lambda E(v) + \bar{f}_v(\lambda). \qquad (8.3.14) \end{aligned}$$

Substitution in (8.3.2) gives

$$Q(z) = P(z)\frac{1 - \tilde{f}_v[\lambda(1-z)] + (1-z)\,\tilde{f}_v(\lambda)}{(1-z)[\lambda E(v) + \tilde{f}_v(\lambda)]}. \tag{8.3.15}$$

8.3.4 Poisson input queue with vacations: waiting-time distribution

By *sojourn time* or *waiting time* in the system of a customer, we shall here mean his queueing time plus his service time. We shall consider steady-state waiting time of an arbitrary customer in a vacation system with Poisson input. We assume the following:

(i) the queue discipline is FIFO (for while queue-length distribution is not affected by queue discipline, waiting-time distribution is affected) and

(ii) the waiting time of a customer is independent of the input process that occurs after the epoch of arrival of the customer considered. Note that this condition is not satisfied in case of a queue under N-policy. Our result will not hold good for such a queue.

Let $W_1(.)$ denote the DF of the waiting time of an arbitrary customer in a standard $M/G/1$ queue (without vacation), and let $W_1^*(.)$ denote its LST. Let $W(.)$ and $W^*(.)$ denote the corresponding functions in an $M/G/1$ queue with vacation under the assumption previously stated. We have the following result due to Fuhrmann and Cooper (1985).

Theorem 8.7. *For an $M/G/1$ queue with vacations,*

$$W^*(s) = W_1^*(s)\,V\left(1 - \frac{s}{\lambda}\right). \tag{8.3.16}$$

Proof: Under FIFO discipline, the customers left behind by an (arbitrary) departing customer are precisely those customers that arrived during the waiting time of the departing customer. It follows that

$$\begin{aligned} Q(z) &= \int_0^\infty \exp\{-\lambda(1-z)t\}\,dW(t) \\ &= W^*[\lambda(1-z)], \end{aligned}$$

so that, putting $\lambda(1-z) = s$, we get

$$W^*(s) = Q\left(1 - \frac{s}{\lambda}\right) \tag{8.3.17}$$

for an $M/G/1$ queue with vacation. Similarly, for the queue without vacation

$$W_1^*(s) = P\left(1 - \frac{s}{\lambda}\right). \tag{8.3.18}$$

Using the basic decomposition result (8.3.1), one gets

$$W^*(s) = W_1^*(s)V\left(1 - \frac{s}{\lambda}\right).$$

Case i. $M/G/1 - V_m$

Putting $s = \lambda(1 - z)$ in (8.3.9), we get

$$Q\left(1 - \frac{s}{\lambda}\right) = P\left(1 - \frac{s}{\lambda}\right)\frac{1 - \bar{f}_v(s)}{sE(v)},$$

so that for such a system

$$W^*(s) = W_1^*(s)\frac{1 - \bar{f}_v(s)}{sE(v)}, \tag{8.3.19}$$

where $W_1^*(s)$ is given by the Pollaczek-Khinchin formula.
Note that

$$V\left(1 - \frac{s}{\lambda}\right) = \frac{1 - \bar{f}_v(s)}{sE(v)}$$

is the LST of the forward-recurrence time of a vacation (as can be observed from an earlier discussion). In this vacation model, the waiting-time distribution decomposes into two independent components. One is the waiting-time distribution in the corresponding model without vacation. The other is the forward-recurrence time of the vacation. When the vacation distribution is given, the decomposition result reduces the problem to a convolution problem.

We have followed Fuhrmann's approach. Alternative methods of derivation of decomposition result have been given by Shanthikumar (1988) through the level-crossing argument and by Doshi (1985) through the sample-path argument.

Mean queueing time $E\{W_Q\}$ in a vacation model is given by

$$\begin{aligned} E\{W_Q\} &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E(V^2)}{2E(V)} \\ &\quad \text{(where } b^{(2)} \text{ is the second moment of service time),} \\ &= \text{mean waiting time in the standard } M/G/1 \text{ queue} \\ &\quad + \text{mean residual vacation time,} \end{aligned}$$

that is,

$$\begin{aligned} &E\{W_Q\}(M/G/1 - V_m) \\ &\quad = E\{W_Q\}(M/G/1) + \text{mean residual vacation time.} \end{aligned}$$

Case ii. $M/G/1 - V_s$ model

Putting $s = \lambda(1 - z)$ in (8.3.15), we get (using (8.3.17) and (8.3.18))

$$W^*(s) = W_1^*(s)\frac{\bar{f}_v(\lambda) + \left(\frac{\lambda}{s}\right)[1 - \bar{f}_v(s)]}{\lambda E(v) + \bar{f}_v(\lambda)}. \tag{8.3.20}$$

■

8.3.5 $M/G/1$ system with vacations: nonexhaustive service

So far we assumed that once the server starts service, he serves all the customers one by one until none is left. As against this exhaustive service discipline, there could be situations giving rise to a nonexhaustive service discipline, under which the server vacation may start even when some customers are present in the system (left in the queue without being served). Two cases may arise.

(i) The preemptive case where vacations may preempt an ongoing service—for example, as in the case of a breakdown of the service mechanism, and

(ii) nonpreemptive case where the vacations may commence only at epochs of service completion or vacation termination—for example, as in the case of scheduled preventive maintenance.

We shall restrict ourselves here to the nonpreemptive case. Assume that service discipline is LIFO. Consider an $M/G/1$ system with vacations. Let R denote the number of customers already present when a typical vacation began and let $\zeta(z)$ be its PGF. Consider an epoch of vacation commencement when the queue length is R. The vacation and the customers who arrive during this vacation will start a cycle that will not be affected by the R customers present at its commencement. Let R be independent of the arrival process during a vacation.

Suppose that we suppress all vacations during this cycle. Then the distribution of the queue length (excluding the R customers initially present) during this period is the same as that of an $M/G/1$ queue with exhaustive service discipline. Thus, the number of customers Q_1 at a departure epoch of an $M/G/1$ vacation queue with nonexhaustive service discipline will be equal to $R + Q$ where Q is the number of customers at a departure epoch of the corresponding $M/G/1$ vacation queue with exhaustive service discipline. Now Z and Q are independent. Denoting the PGF of Q_1 by $Q_1(z)$, we shall get the following three-way decomposition result.

Theorem 8.8. *For an $M/G/1 - V_m$ queue with nonexhaustive service, the PGF $Q_1(z)$ of the number at a departure epoch is given by*

$$\begin{aligned} Q_1(z) &= \zeta(z)\,Q(z) \\ &= \zeta(z)\,P(z)\frac{1 - \alpha(z)}{(1 - z)\alpha'(1)}. \end{aligned} \tag{8.3.21}$$

For a rigorous proof and observations on this interesting decomposition result, refer to Fuhrmann and Cooper (1985) *and Doshi* (1986).

Notes:

(1) The preceding three-way decomposition result will hold if the number R present at the commencement of a vacation and the number of arrivals A_v during a vacation are independent.

(2) For exhaustive service, $Pr(R = 0) = 1$ and $\zeta(z) = 1$; then (8.3.21) reduces to (8.3.12).

(3) For nonexhaustive service

$$Q(z) = \chi(z)P(z),$$

where

$$\chi(z) = \frac{\zeta(z)(1 - \alpha(z))}{(1 - z)\alpha'(1)}$$

is the PGF of the number of units at a random point of time in the vacation period.

8.3.6 Limited service system: $M/G/1 - V_m$ model

There are specific types of nonexhaustive service, such as Limited Service, Gated Service, Decrementing Service, and so on. (See Takagi (1991) for a full description and discussion.)

Consider a multiple vacation system. In the limited service system, the server takes a vacation each time it completes service of a unit. When he returns from a vacation, it serves others remaining in the queue, if any, and if there is none, then it takes another vacation, and so on until there is at least one unit in the system. This is called a (pure) limited service system.

Replacing $B^*(s)$ by $B^*(s)V^*(s)$ in (8.3.19), we get, for such a limited service system, the LST of the waiting time:

$$W^*(s) = \frac{s(1 - \rho - \lambda E(V))}{s - \lambda[1 - B^*(s)V^*(s)]} \cdot \frac{1 - V^*(s)}{sE(V)},$$

where v is service time and V is vacation time.

Now, $E(v)$ becomes $E(v) + E(V)$

The mean waiting time is given by

$$E\{W_Q\} = \frac{\lambda\{b^{(2)} + 2E(v)E(V) + E(V^2)\}}{2(1 - \rho - \lambda E(V))} + \frac{E(V^2)}{2E(V)}.$$

The mean response time is given by $E\{W\} = E\{W_Q\} + b$

Limited Service with Bernoulli Scheduled Vacation

Here, after completion of service, the server either takes a vacation of duration V with probability $(1 - \tau)$ or continues to serve the next unit, if any, with probability τ $(0 \leq \tau \leq 1)$.

In this case $B^*(s)$ is to be replaced by

$$\tau B^*(s) + (1 - \tau) B^*(s) V^*(s),$$

and $E(v)$ by

$$\tau b + (1 - \tau)\{b + E(V)\} = b + (1 - \tau) E(V)$$

and the utilization factor by

$$\lambda b + (1 - \tau)\lambda E(V)$$

Thus,

$$W^*(s) = \frac{s[1 - \rho - (1 - \tau)\lambda E(V)]}{s - \lambda + \lambda[\tau B^*(s) + (1 - \tau) B^*(s) V^*(s)]} \cdot \frac{1 - V^*(s)}{s E(V)}.$$

The mean waiting time is given by

$$E\{W\} = \frac{\lambda\left[b^{(2)} + (1 - \tau)\{2bE(V) + E(V^2)\}\right]}{2[1 - \rho - (1 - \tau)\lambda E(V)]} + \frac{E(V^2)}{2E(V)}.$$

For $\tau = 1$, one gets an exhaustive service system, and for $\tau = 0$, one gets a limited service system.

Such models could be used in computer and communication systems.

For applications to computer and communication systems and other types of limited service systems, see Takagi (1991).

8.3.7 Gated service system: $M/G/1 - V_m$ model

Consider a nonexhaustive service system such that the server, when he returns from a vacation, finds a random number R waiting for service. The server then accepts service only those R units and arrivals after the server starts service during his service-period remains in the queue without being served. The server leaves for a vacation soon after serving the R units only (as if the R waiting units when the server arrives are brought inside a gate and the gate is closed and others who arrive after the server starts service—during his service period—queue up outside the gate to be served after the server's vacation).

In the multiple vacation model, the server goes for another vacation till he finds at least one unit waiting, when he begins serving. There are other types of gated service systems as well. Gated service is one particular type of nonexhaustive service.

We can now apply Theorem 8.8 above for this kind of gated service system. The PGF of the number in the system is given by (8.3.21)

$$Q_1(z) = \zeta(z)\, Q(z),$$

where $\zeta(z)$ = PGF of the number L present at the beginning of a typical vacation
= PGF of the number L of arrivals during the service period (busy period) of the server,

and

$Q(z)$ = PGF of the number left behind by a departing customer of an $M/G/1 - V_m$ queue with exhaustive service discipline.

Let S be defined as the length of the service period (the period between two vacations). (S is the sojourn period of the server.) Note that $Pr\{S = 0\} \neq 0$.

If $S^*(.)$ is the LST of S, then

$$\begin{aligned} \zeta(z) &= \text{PGF of the number of arrivals } A_1 \text{ during the period } S \\ &= S^*(\lambda - \lambda z). \end{aligned} \tag{8.3.22}$$

Now

$$\begin{aligned} h(z) &= \text{PGF of the number of arrivals } A_2 \text{ during a vacation} \\ &= V^*(\lambda - \lambda z). \end{aligned} \tag{8.3.23}$$

The number served during a service period is the sum of two independent RVs A_1 and A_2. Thus the PGF $A(z)$ of the number α served during a service period is given by

$$A(z) = S^*(\lambda - \lambda z)\, V^*(\lambda - \lambda z). \tag{8.3.24}$$

Now the total length of S is the sum of the service times of a random number of customers with PGF $A(z)$. The LST $S^*(s)$ of the service times of these α customers is thus

$$S^*(s) = A(B^*(s)). \tag{8.3.25}$$

Thus, from (8.3.24) and (8.3.25) one gets

$$A(z) = A[B^*(\lambda - \lambda z)]\, V^*(\lambda - \lambda z). \tag{8.3.26}$$

From the above, one gets

$$E(\alpha) = \lambda E(\alpha) b + \lambda E(V)$$

$$\text{or} \quad E(\alpha) = \frac{\lambda E(V)}{1 - \rho}, \tag{8.3.27}$$

which is the expected number in the queue at the beginning of a vacation (= number of arrivals during a typical service period). Further,

$$E(S) = bE(\alpha) = \frac{\rho E(V)}{1-\rho}, \tag{8.3.28}$$

and the expected number of arrivals L during an expected busy period is

$$E(L) = \lambda E(S) = \frac{\lambda\rho E(V)}{1-\rho}$$
$$\text{also} \quad E(L) = -\xi'(z)|_{z=0}. \tag{8.3.29}$$

From the relation

$$Q_1(z) = \zeta(z)\, Q(z),$$

we find the expected number N_1 of units in a gated service system as

$$E\{N_1\} = E\{N\} + E\{L\},$$

where N = number in the system in an $M/G/1 - V_m$ queue with exhaustive service.

Thus,

$$E\{N_1\} = \rho + \frac{\lambda^2 b^{(2)}}{2(1-\rho)} + \frac{\lambda E(V^2)}{2E(V)} + \frac{\lambda\rho E(V)}{1-\rho}. \tag{8.3.30}$$

Using Little's formula, one gets

$$E\{W_1\} = E\{W\} + E\{L\}/\lambda = E\{W\} + E\{S\}. \tag{8.3.31}$$

Thus,

$$E\{W_1\}_{\text{gated}} = E\{W\}_{\text{exhaustive}} + E\{S\}. \tag{8.3.32}$$

Using (8.3.22), (8.3.25), and (8.3.26), one gets from (8.3.23)

$$Q_1(z) = A(B^*(\lambda - \lambda z))\, P(z) \cdot \frac{1 - V^*(\lambda - \lambda z)}{\lambda(1-z)E(V)} \tag{8.3.33}$$

$$= A(B^*(\lambda - \lambda z)) \frac{(1-\rho)(1-z)\, B^*(\lambda - \lambda z)}{B^*(\lambda - \lambda z) - z} \cdot \frac{1 - V^*(\lambda - \lambda z)}{\lambda(1-z)E(V)}$$

$$= \frac{[A(B^*(\lambda - \lambda z)) - A(z)][B^*(\lambda - \lambda z)]}{E(\alpha)[B^*(\lambda - \lambda z) - z]}, \tag{8.3.34}$$

which is the PGF of the number in the system in an $M/G/1 - V_m$ queue with gated service system.

Notes:

(1) Assuming that (8.3.23) holds, we obtain (8.3.34). It is possible to obtain (8.3.34) directly and then get the three-way decomposition (8.3.33). (See Takagi (1991).)

(2) The RV S is the time taken to serve $\alpha = \alpha_1 + \alpha_2$ units and is therefore the same whether the service is exhaustive or gated.

Busy and Idle Periods: T and I
For an $M/G/1$ queue,

$$E(T) = \frac{1}{\mu(1-a)}, \qquad E(I) = \frac{1}{\lambda}, \quad a = \rho,$$

so that

$$E(T) = \frac{1}{\mu - \lambda} = \frac{\rho}{1-\rho} E(I).$$

This relation also holds for other modifications of single-server queues with Poisson input.

Consider an $M/G/1 - V_m$ queue. Here

$$V^*(\lambda) = \int_0^\infty e^{-\lambda t}\, dV(t) = Pr\{\text{None arrives during a vacation}\}$$

$$\text{and} \quad 1 - V^*(\lambda) = Pr\{\text{at least one arrival during a vacation}\}.$$

Thus, I is the sum of a random number of IID random variables, the sum being a geometric RV with mean $\frac{1}{1-V^*(\lambda)}$. Thus,

$$E\{I\} = \frac{E(V)}{1 - V^*(\lambda)}, \tag{8.3.35}$$

and the average number of arrivals during $E(I)$ is $\lambda E(I)$. That is, a busy period starts with an average number of $\lambda E(I)$ units. Hence, the length of an average busy period is given by

$$E\{T\} = \frac{\lambda E(I)}{\mu - \lambda}.$$

Consider an $M/G/1 - V_s$ system. Conditioning on where during a vacation, there is at least one arrival or none at all, one gets

$$E(I) = (1 - V^*(\lambda))E(V) + V^*(\lambda)\left[E(V) + \frac{1}{\lambda}\right]$$

$$= E(V) + \frac{V^*(\lambda)}{\lambda}. \tag{8.3.36}$$

Thus,

$$E(T) = \frac{\lambda E(I)}{\mu - \lambda} = \frac{\lambda E(V) + V^*(\lambda)}{\mu - \lambda}. \tag{8.3.37}$$

Assume that: The waiting time of a customer is independent of the arrival process that occurs after the epoch of customer's arrival. This assumption is satisfied in vacations with a FCFS queueing discipline (but not for N-policy).

Theorem 8.9. *Let $W_1^*(.)$ and $W^*(.)$ denote the LST and $W_1(.), W(.)$ the DF, respectively, of the waiting time of the $M/G/1$ vacation model and the corresponding standard $M/G/1$ model. Then*

$$W_1^*(s) = F^*(s)W^*(s), \tag{8.3.38}$$

where $F^*(s)$ the product of the LST of the service period and the LST of the residual vacation.

Proof: Under FCFS discipline when the above assumption holds, the units left behind by a departing unit are those that arrived during the unit's waiting time, and

$$\begin{aligned} Q_1(z) &= \int_0^\infty e^{-\lambda t(1-z)} dW_1(t) \\ &= W_1^*(\lambda - \lambda z). \end{aligned}$$

From (8.3.33)

$$\begin{aligned} W_1^*(s) = Q_1\left(1 - \frac{s}{\lambda}\right) &= P\left(1 - \frac{s}{\lambda}\right) A(B^*(s)) \frac{1 - V^*(s)}{\lambda \cdot \frac{s}{\lambda} E(V)} \\ &= W^*(s)S^*(s) \cdot \frac{1 - V^*(s)}{sE(V)} \\ &= W^*(s)F^*(s). \end{aligned} \tag{8.3.39}$$

where

$$F^*(s) = [S^*(s)]\left[\frac{1 - V^*(s)}{sE(V)}\right]. \tag{8.3.40}$$

That is, $F^*(s)$ is the product of the LST of the service period (sojourn period) and the LST of the residual vacation. ■

8.3.8 $M/G/1/K$ queue with multiple vacations

An $M/G/1/K$ queue with multiple vacation and exhaustive service in steady state has been considered by Lee (1984) and Frey and Takahashi (1997a,b). While Lee considers both service completion and vacation completion epochs,

the latter consider only service completion epochs. We follow this latter approach here. K is the number of waiting places, including the place for the customer in service, if any. Arrivals when all the place are full are lost to the system.

Denote by

$$g_j = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} dB(t), \quad j = 0, 1, 2, \ldots \tag{8.3.41}$$

$$h_j = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} dV(t), \quad j = 0, 1, 2, \ldots, \tag{8.3.42}$$

the probability that j arrive during a service time (DF $B(t)$) of a unit and during a vacation time V(df $V(t)$) of the server, respectively. The probabilities that j units arrive and are accepted when the server is on vacation are given by

$$\begin{aligned} \varphi_j &= \frac{h_j}{1 - h_0}, \quad j = 0, 1, \ldots, K-1 \\ &= \frac{\sum_{j=K}^\infty h_j}{1 - h_0}, \quad j = K, \end{aligned} \tag{8.3.43}$$

where $h_0 = V^*(\lambda)$.

Denote by π_j, $j = 0, 1, \ldots, K-1$, the probability that j units are left in the system at a departure epoch of a unit.

Then π_j satisfy the following equations

$$\pi_j = \pi_0 \sum_{i=0}^{j+1} \varphi_i g_{j-i+1} + \sum_{i=1}^{j+1} \pi_i g_{j-i+1}, \quad j = 0, 1, \ldots, K-2 \tag{8.3.44}$$

$$\pi_{K-1} = \pi_0 \sum_{i=1}^{K} \varphi_i \left(g_{K-1}^c\right) + \sum_{i=1}^{K-1} \pi_i g_{K-i}^c, \tag{8.3.45}$$

where $g_k^c = \sum_{i=k}^\infty g_i$ with the normalizing condition $\sum_{i=0}^{K-1} \pi_i = 1$.

From the above recursive relations we get $\pi_1, \pi_2, \ldots, \pi_{K-1}$ in terms of π_0, and using $\sum_{i=0}^{K-1} \pi_i = 1$, we can evaluate all the π_i's, $j = 0, 1, \ldots, K-1$.

See Frey and Takahashi (1997b) for exact expressions of $\{\pi_j, j = 0, 1, \ldots, K-1\}$.

Frey and Takahashi (1997a,b) also obtain the probabilities π_j^* that there are j in the system at an arbitrary time in steady state: These are given by

$$\pi_j^* = \frac{\pi_j (1 - V^*(\lambda))/\lambda}{\pi_0 E(V) + E(B)(1 - V^*(\lambda))}, \quad j = 0, 1, \ldots, K-1 \tag{8.3.46}$$

$$\pi_K^* = 1 - \frac{(1 - V^*(\lambda))/\lambda}{\pi_0 E(V) + E(B)(1 - V^*(\lambda))}. \tag{8.3.47}$$

They obtain the LST of the steady-state waiting-time distribution.

Further, by taking the vacation time to be a constant V and letting $V \to 0$, they obtain the steady-state distribution of the number in the system at an arbitrary time for the $M/G/1/K$ queue.

8.3.8.1 $G/G/1$ model

An analogous type of decomposition result involving waiting time holds for a $GI/G/1 - V_m$ model. This has been demonstrated, for example, by Gelenbe and Iasnogorodski (1980), Doshi (1985, 1986), and Fricker (1986). The sample-path arguments put forward by Doshi (1985) could be extended to cover $GI/G/1 - V_m$ and $GI/G/1 - V_s$ models. For a $GI/G/1 - V_m$ system,

$$E(W) = E(W)(GI/G/1) + \text{mean residual vacation time.}$$

The decomposition result holds. There has been further study in this direction for more general (arrival pattern) models. Lucantoni *et al.* (1990) show that, for single-server queues with exhaustive service and multiple vacations, decomposition of this type holds for a class of nonrenewal arrival process called MAP (Markovian arrival process). See also Doshi (1990).

8.3.8.2 Variations of the vacation model

There are various related models—for example, a model with start-up time. Here, when a customer arrives to start a busy period, the server goes through a set-up or start-up time (SUT) of random length U before starting actual service. The server is unavailable to primary customers during this start-up time.

This problem and its variations have been considered by several researchers, for example, Pakes (1972), Lemoine (1975), and Doshi (1985).

For a Poisson input queue with SUT U, with DF $F_u(.)$, it is shown that the decomposition can be put as

$$Q(z) = P(z)\frac{1 - R(z)}{(1 - z)\lambda E(U)}, \tag{8.3.48}$$

where

$$\begin{aligned} R(z) &= 1 \times z + \sum_{n=0}^{\infty} z^n \int_0^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} dF_u(t) \\ &= z + \tilde{f}_u[\lambda(1 - z)] \end{aligned} \tag{8.3.49}$$

$\tilde{f}_u(.)$ being the LST of $F_u(.)$.

The decomposition property has been used by Medhi and Templeton (1992) in the study of an $M/G/1$ queue under N-policy, with general start-up time. Fuhrmann and Cooper (1985) have considered application of decomposition property in an $M/G/1$ queue with vacations to two continuum cyclic queueing models.

A finite-capacity vacation-type queue has been considered, for example, by Teghem (1987) and Loris-Teghem (1988), for an $M/G/1$ model, and by Jacob and Madhusoodanan (1987) for an $M/G(a,b)/1$ model.

Notes:

(1) The time-dependent process of $M/G/1$ vacation models with exhaustive service has been discussed by Keilson and Servi (1987), Ramaswamy and Keilson (1988), and Takagi (1990, 1992b), for both multiple and single vacations.

(2) Takagi (1992) examines a finite population $M/G/1//N$ model with multiple vacations and exhaustive service and its application to a polling model, where he obtains the expressions of several performance measures, including the throughput, mean response time, mean idle, and busy periods.

(3) Bulk arrival $M^X/G/1$ queues with vacations have been considered, among others, by Baba (1986), Altiok (1987), Choudhury (2000, 2001), and Frey and Takahashi (2000). Such queues under N-policy have been studied by Lee *et al.* (1994) for multiple vacations and by Lee *et al.* (1995) for single vacations.

$GI/M/1$ system with vacations

The $GI/M/1$ queue with vacations has been studied by Tian *et al.* (1989) and Chatterjee and Mukherjee (1990). They consider the infinite waiting space case. Karaesmen and Gupta (1996) examine the $GI/M/1/K$ queue (finite waiting space) with multiple vacations and exhaustive service-vacation durations having exponential distribution. They consider the Markov chain embedded at arrival epochs and obtain the equations for the limiting probabilities of the embedded Markov chain; the solution of the equations could be obtained numerically. Tian *et al.* (1989) obtain closed-form expressions in the case of infinite waiting space system.

Introduction of a vacation component makes many queueing systems more realistic. It can also be visualized that many queueing problems can be simplified by considering them as vacation-type problems, and their solution too can be simplified.

8.3.8.3 Multiserver models

It appears that relatively little attention has been paid to multiserver queues with vacations. Multiserver queues $M/M/c$ with vacations have been considered, for example, by Mitrani and Avi-Itzhak (1968), Levy and Yechiali (1976), and Neuts and Lucantani (1979). The queues with general service time do not seem to have been studied.

Remark: It may be observed in passing that there are similar, though not analogous, product-form results of interest in other areas. For example, we may mention the product-form or factorization result of sufficient statistics in

the theory of statistical estimation. Another interesting result is the "product-form" representation of the conditional hazard function $h(t \mid X, \theta)$ on a vector of covariated X and unobserved heterogeneity component θ. It is as follows.

$$h(t \mid X, \theta) = h_0(t)\psi(X)\phi(\theta),$$

where $h_0(t)$ is the baseline hazard function (Cox, 1972). This representation has been fruitfully employed in several types of studies, such as in econometrics (for example, Lancaster (1979) and Heckman and Singer (1984)) and in business in the study of household-brand-switching behavior.

8.3.9 Mean value analysis through heuristic treatment

Chae and Lee (1995) give a heuristic treatment of the mean waiting time of a number of queueing models with vacation. It could be applied to a large number of other models as well, and a unified treatment can be brought about for the mean waiting time (see also Medhi (1997)). Mean value analysis of performance measures is useful for many practical purposes. Once mean waiting time is found, mean system size follows.

Denote

W_Q = mean waiting time (queueing time)
L_Q = mean member in queue
$a = \lambda E(B)$ (offered load), B = service time
ρ = server utilization factor
I = length of (server) idle period
T = length of busy period

8.3.9.1 M/G/1 queue

The mean queueing time (mean waiting time in the queue) is easily seen to be

$$W_Q = L_Q E(B) + E(B_R) \cdot Pr\{\text{server is busy}\}, \tag{8.3.50}$$

where B_R is the residual service time. We have $E(B_R) = \frac{b^{(2)}}{2b}$, where $b, b^{(2)}$ are the first two moments of the service time. Also, $L_Q = \lambda W_Q$. Thus,

$$W_Q = \lambda E(B) W_Q + \rho \cdot \frac{b^{(2)}}{2b},$$

whence

$$W_Q = \frac{\lambda b^{(2)}}{1-\rho}, \quad \rho = a(= \lambda E(B)). \tag{8.3.51}$$

Also,

$$E(T) = \frac{\rho}{1-\rho} E(I).$$

8.3.9.2 *M/G/1 with exceptional service for the first unit within each busy period*

Assume that the first unit in a busy period has service time B_0 with first two moments $b_0, b_0^{(2)}$, respectively, and other units have service time B with first two moments $b, b^{(2)}$, respectively. Here

$$E(I) = 1/\lambda$$

$$E(T) = b_0 + \frac{(\lambda b_0) b}{1-a} = \frac{b_0}{1-a} \tag{8.3.52}$$

$$\text{and} \quad \rho = \frac{E(T)}{E(I) + E(T)} = \frac{\lambda b_0}{1 - a + \lambda b_0}, \tag{8.3.53}$$

so that

$$P_0 = 1 - \rho = \frac{1-a}{1 - a + \lambda b_0}.$$

(Note that $E(T) = \frac{\rho}{1-\rho} E(I)$ holds.)

Denote

B_1 = service time of a unit
= B_0 or B according as the unit is the first to be served in a busy period or later

Thus,

$$E(B_1) = b_0 P_0 + b(1 - P_0)$$

$$= \frac{\rho}{\lambda}, \tag{8.3.54}$$

$$\text{so that} \quad \rho = \lambda E(B_1).$$

$$\text{Again,} \quad E\left(B_1^2\right) = b_0^{(2)} P_0 + b^{(2)}(1 - P_0)$$

$$= \frac{(1-a) b_0^{(2)} + \lambda b_0 b^{(2)}}{1 - a + \lambda b_0}. \tag{8.3.55}$$

$$\text{We have} \quad W_Q = L_Q E(B) + \rho E(B_{1R}), \tag{8.3.56}$$

where $B_{1R} \equiv$ residual service time (B_1) of a unit $= \frac{E(B_1^2)}{2E(B_1)}$. Thus, from the above relations we get

$$W_Q = \frac{\lambda b^{(2)}}{2(1-a)} + \frac{\lambda\left(b_0^{(2)} - b^{(2)}\right)}{2(1-a+\lambda b_0)}. \tag{8.3.57}$$

Note that when $B_0 \equiv B$, the second term vanishes.

8.3.9.3 M/G/1 queue with random setup time

Suppose that at the end of an idle period, the system has to gear up the service mechanism immediately before start of service in a busy period.

That is, the first unit in a busy period needs a setup time S (SUT) of random duration before service is started. S is independent of other variables. Thus,

$$\begin{aligned} E(I) &= \frac{1}{\lambda} + E(S) \\ E(T) &= \frac{\lambda E(I) b}{1-a} = \frac{b + aE(S)}{1-a} \\ \text{and} \quad \rho &= \frac{E(T)}{E(T)+E(I)} = a, \end{aligned} \tag{8.3.58}$$

so that the utilization factor is not affected by the setup time. Also $E(T) = (\frac{\rho}{1-\rho})E(I)$ holds.

Now to find $P_0 =$ probability {system is empty}, we consider the busy cycle D, which is the duration of the period commencing on the arrival of the unit after an empty period L (period during which there is none in the system) to the instant when the system becomes empty again. Thus, $E(L) = 1/\lambda$ and

$$\begin{aligned} E(D) &= [b + E(S)] + \frac{b}{1-a}[\lambda(b + E(S))] \\ &= \frac{1}{1-a}[b + E(S)] \end{aligned} \tag{8.3.59}$$

$$\text{and} \quad P_0 = \frac{E(L)}{E(L)+E(D)} = \frac{1-a}{1+\lambda E(S)}. \tag{8.3.60}$$

We have

$$\begin{aligned} W_Q = {} & L_Q E(B) + E\{\text{Residual service time} - \text{server is busy}\} \\ & Pr\{\text{server is busy}\} + [E\{\text{Setup time} - \text{test unit is the} \\ & \text{first to arrive in the idle period}\}\ Pr\{A \mid C\} \\ & + E\{\text{residual setup time} - \text{test unit arrives during setup time}\} \\ & \times Pr\{B \mid C\}] \times Pr\{\text{server is idle}\} \end{aligned} \tag{8.3.61}$$

where

$A \equiv$ event that the test unit is the first to arrive in the idle period (at the close of the empty period)

$B \equiv$ event that the test unit arrives during the setup time of the idle period

and $C \equiv$ event that the service is idle

$$Pr\{A \mid C\} = \frac{1/\lambda}{E(I)} = \frac{1}{1+\lambda E(S)} \tag{8.3.62}$$

$$Pr\{B \mid C\} = \frac{E(S)}{E(I)} = \frac{\lambda E(S)}{1+\lambda E(S)}. \tag{8.3.63}$$

Note that

$$Pr\{B \mid C\} = 1 - Pr\{A \mid C\}.$$

Substituting the above values in (8.3.61) and using $L_Q = \lambda W_Q$, we get

$$(1-a)W_Q = \rho \cdot \frac{b^{(2)}}{2b} + (1-\rho)\left[\frac{E(S)}{1+\lambda E(S)} + \frac{E(S^2)}{2E(S)} \cdot \frac{\lambda E(S)}{1+\lambda E(S)}\right],$$

so that

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{2E(S)+\lambda E(S^2)}{2\{1+\lambda E(S)\}}. \tag{8.3.64}$$

8.3.9.4 *M/G/1 queue under N-policy (threshold policy)*

Such a queue was first considered by Yadin and Naor (1963). The server in this case waits for the arrival of a preassigned number $N(\geq 1)$ units (called threshold) and then commences service and continues until none is left in the system. We have

$$E(I) = \frac{N}{\lambda}, \qquad E(T) = \frac{Nb}{1-a},$$

The fraction of time the server is busy is given by

$$\rho = \frac{E(T)}{E(I)+E(T)} = \lambda b = a,$$

which is independent of N; also $E(T) = \frac{\rho}{1-\rho}E(I)$.

The expected waiting time is given by

$$W_Q = L_Q E(B) + E(B_R) \cdot Pr\{\text{server is busy}\} + E(I_R) \cdot Pr\{\text{server is idle}\}, \tag{8.3.65}$$

where B_R = residual service time, I_R = residual idle period (that is, the duration from the instant of the arrival of the test customer in the idle period to the instant of arrival of the Nth unit).

Conditioning on the order in which the test customer arrives, we have

$$E(I_R) = \left(\frac{N-1}{\lambda} + \frac{N-2}{\lambda} + \cdots + \frac{1}{\lambda} + 0\right) \cdot \frac{1}{N} = \frac{N-1}{2\lambda}. \tag{8.3.66}$$

Thus, from (8.3.65), we get

$$\begin{aligned} (1-a)W_Q &= \rho \cdot \frac{b^{(2)}}{2b} + (1-\rho) \cdot \frac{N-1}{2\lambda} \\ \text{or} \quad W_Q &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{N-1}{2\lambda}. \end{aligned} \tag{8.3.67}$$

8.3.9.5 M/G/1 queue under N-policy and with setup time

We consider the situation where the server, as soon as the queue builds up to N (after the idle period), needs a setup time of random duration S before starting service exhaustively until none is left in the system.

The idle period is comprised of the buildup period (duration of time for arrival of Nth unit after the system becomes empty) plus the setup period S. Thus,

$$E(I) = \frac{N}{\lambda} + E(S)$$

$$E(T) = \lambda E(I) \cdot \frac{b}{1-a} = \frac{a}{1-a}\left(\frac{N}{\lambda} + E(S)\right).$$

$$\text{Further} \quad \rho = \frac{E(T)}{E(I) + E(T)} = a,$$

independent of the buildup and the setup period. This is also shown by Medhi and Templeton (1992). The expected waiting time is given by

$$\begin{aligned} W_Q = {} & L_Q E(B) + E(\text{Residual service time} \mid \text{server is busy}) \\ & \times Pr\{\text{server is busy}\} + [E\{\text{Residual buildup period} \mid A\}\, Pr\{A \mid C\} \\ & + E\{\text{Residual setup period} \mid B\}\, Pr\{B \mid C\}] \times Pr\{\text{server is idle}\}, \end{aligned} \tag{8.3.68}$$

where

A = event that the test customer arrives in the buildup period

B = event that the test customer arrives in the setup period

C = event that the server is idle

$$\text{Now} \quad Pr\{A \mid C\} = \frac{E\{\text{length of buildup period}\}}{E\{\text{length of idle period}\}}$$

$$= \frac{N/\lambda}{N/\lambda + E(S)} = \frac{N}{N + \lambda E(S)}$$

$$Pr\{B \mid C\} = \frac{E\{\text{length of setup period}\}}{E\{\text{length of idle period}\}}$$

$$= \frac{E(S)}{N/\lambda + E(S)} = \frac{\lambda E(S)}{N + \lambda E(S)}$$

Thus, from (8.3.68) we get

$$(1-a)W_Q = \rho\frac{b^{(2)}}{2b} + (1-\rho)\left[\left\{\frac{N-1}{2\lambda} + E(S)\right\} \cdot \frac{N}{N+\lambda E(S)} + \frac{E(S^2)}{2E(S)} \cdot \frac{\lambda E(S)}{N+\lambda E(S)}\right],$$

whence

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{N(N-1) + 2\lambda N E(S) + \lambda^2 E(S^2)}{2\lambda[N + \lambda E(S)]}. \tag{8.3.69}$$

(Yadin and Naor, 1963; Takagi, 1991)

8.3.9.6 $M/G/1 - V_m$ queue

The idle period I is the duration from the instant a busy period ends and the server takes repeated vacations up to the instant that a unit arrives and service begins.

If $V(t)$ is the DF of a vacation, then

$$V^*(\lambda) = \int_0^\infty e^{-\lambda t} dV(t)$$

$$= Pr\{\text{none arrives during a vacation}\}.$$

The idle period I is the sum of a random number α of vacations, each of length V; α is a geometric r.v. with probability of success $\{1 - V^*(\lambda)\}$. Thus,

$$E(I) = \frac{E(V)}{1 - V^*(\lambda)}, \tag{8.3.70}$$

so that

$$E(T) = \{\lambda E(I)\} \cdot \frac{b}{1-a} = \frac{aE(V)}{(1-a)(1-V^*(\lambda))}$$

and

$$\rho = \frac{E(T)}{E(I)+E(T)} = a;$$

also

$$E(T) = \frac{\rho}{1-\rho} E(I).$$

The expected waiting time is given by

$$\begin{aligned} W_Q = {} & L_Q E(B) + E(B_R \mid \text{server is busy}) \cdot Pr\{\text{server is busy}\} \\ & + E\{V_R \mid \text{server is on vacation}\} \cdot Pr\{\text{server is idle}\} \\ & (B_R = \text{residual service time}; V_R = \text{residual vacation time}) \end{aligned} \tag{8.3.71}$$

Thus,

$$\begin{aligned} W_Q &= \frac{1}{1-a} [\rho E(B_R) + (1-\rho) E(V_R)] \\ &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E(V^2)}{2E(V)}. \end{aligned} \tag{8.3.72}$$

8.3.9.7 $M^X/G/1$ system

Suppose that arrivals occur in batches of random size X, the instants of arrivals being Poissonian with rate λ, and service time S is general with mean $b(=1/\mu)$.

Here $\rho = \lambda E(X) b = Pr\{\text{server is busy}\}$; assume that $\rho < 1$.

The mean queueing time

$$W_Q = L_Q \cdot b + E(X_R) b + \rho E\{\text{residual service time}\}, \tag{8.3.73}$$

where X_R = number of units of the group in which the test unit arrived and are served prior to the test unit. The expected delay that a test unit will have due to this group equals

$$E(X_R) \cdot b = E(D_2) = \frac{1}{2}\left[\frac{E(X^2)}{E(X)} - 1\right] \cdot b$$

(see Eq. (6.7.12), Ch. 6).

Thus, we have

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{1}{2(1-\rho)}\left[\frac{E(X^2)}{E(X)} - 1\right] \tag{8.3.74}$$

$M^X/G/1$ with vacations

Consider first the case of multiple vacations—that is, $M^X/G/1 - V_m$ Queue. Here

$$W_Q = L_Q \cdot b + E(B_R) \cdot \rho + E(X_R) \cdot b + E(V_R)(1-\rho),$$

whence

$$W_Q = \frac{\lambda E(X) b^{(2)}}{2(1-\rho)} + \frac{b}{2(1-\rho)}\left\{\frac{E(X^2)}{E(X)} - 1\right\} + \frac{E(V^2)}{2E(V)}. \tag{8.3.75}$$

Consider then the case of a single vacation—that is, $M^X/G/1 - V_s$ Queue.

Here

$$E(I) = E(V) + V^*(\lambda) \cdot \frac{1}{\lambda}$$

and Pr\{server is on vacation | server is idle\}

$$= \frac{E(V)}{E(I)} = \frac{\lambda E(V)}{\lambda E(V) + V^*(\lambda)}.$$

Thus,

$$W_Q = \frac{\lambda E(X) b^{(2)}}{2(1-\rho)} + \frac{b}{2(1-\rho)}\left\{\frac{E(X^2)}{E(X)} - 1\right\} + \frac{E(V^2)}{2E(I)}. \tag{8.3.76}$$

8.4 Design and Control of Queues

Our main concern so far has been to describe probabilistically the behavior of a system and to find the various associated performance measures, given the arrival and service patterns, queue discipline, and other configurations. Practical questions that arise in applications of queueing theory relate to determination of parameters, patterns, and/or policies for which a model will be optimal in some specified sense. In general, one may have little or no control over the arrival patterns, though sometimes arrivals can be controlled through truncation or other means. On the other hand, it may be possible to exercise control over the service mechanism (for example, through the service pattern or number of service channels to be operated), the service policy, and such other configurations in a queueing system in order that the system performance may attain optimal value in a certain sense. For example, control may

be through adjustment of service rate (Mitchell, 1973; Doshi, 1978) or by turning on and off a service mechanism according to a policy involving the state of the system (Heyman, 1968). Objective functions as well as constraints can be formulated in terms of such measures as average cost, average waiting time, and other performance measures of interest. Solutions need optimization techniques. Problems that arise in such situations are called problems of design and control of queues. It is, however, not always easy to make a clear distinction between design and control problems and their corresponding models.

Classical optimization methods are generally used in designs. Problems, formulations, and analyses of control problems involve application of theory and techniques of the renewal reward process, Markov decision processes, the martingale, dynamic programming, and so on.

Optimization models have been increasingly used in the design problems arising in several applied areas, such as production processes and telecommunication networks. Wherever queueing phenomena arise in such problems, performance measures of associated queueing systems often occur in the objective functions and/or in the constraints of the optimization models. The solution of certain optimization problems depends partly on the *concavity* and *convexity properties* of these measures. As such, investigation of these properties has been receiving increasing attention in recent times—for example, see Rolfe (1971), Dyer and Proll (1977), Tu and Kumun (1983), Weber (1983), Grassman (1983), Lee and Cohen (1983), and Harel and Zipkin (1987).

The design problem associated with server allocation has several applications in such areas as multiple center manufacturing systems (see, for example, Shanthikumar and Yao (1986, 1987, 1988)), as well as allocation of vehicles and fleets in transportation, police patrol and ambulances, and so on (for example, Parikh (1977), Green (1984), Berman *et al.* (1985), Chaiken and Dermont (1978), and so on).

A detailed discussion of the topic (design and control of queues) is beyond the scope of this book. (For a detailed account, refer to Kitaev and Rykov (1995).) See also Kushner (2001).

We would, however, try to indicate briefly the nature of the problems that arise. For a general survey of the topics, see, for example, Stidham and Prabhu (1974), Crabill *et al.* (1977), Sobel (1974), and Serfozo (1981). Hillier and Lieberman (1967) described a number of interesting design problems. One simple design model considered is as follows.

Consider a single-server model with known λ. Assume that the cost per server per unit time is A and that the cost of waiting per customer per unit time is B. The design problem consists of finding μ that will minimize the expected cost per unit time.

$$E(C) = A\mu + BL,$$

where L is the average number of customers in the system.

For an $M/M/1$ queue

$$E(C) = A\mu + \frac{B\lambda}{(\mu - \lambda)},$$

so that the value of μ for which $E(C)$ is minimum is the value μ^*, if any, that satisfies

$$\frac{d}{dC}E(C) = 0 \quad \text{and}$$
$$\frac{d^2}{dC^2}E(C) < 0;$$

we get

$$\mu^* = \lambda + \sqrt{\frac{\lambda B}{A}}.$$

It may be noted that even for an $M/M/c$ model, $E(C)$ is at a minimum for $c = 1$ and so for the same μ^*.

Consider now a control problem. One control policy that consists of turning the server on and off is the N-policy formulated by Heyman (1968) (and discussed in the previous section and Section 6.4.5). The facility is shut down (tuned off) as soon as all present are served and reactivated (turned on) as soon as the queue size, when the server is idle, builds up to N. N is called the control parameter. Heyman considers (i) a start-up cost, (ii) a shut-down cost, (iii) a server running cost per unit time (when the server is active), and (iv) a customer-holding (-waiting) cost per customer per unit time. He shows that the optimal type of policy is the N-policy.

Similarly, T-policy is another control policy, according to which the service facility is turned off for a fixed period of time, T, from the instant of each service completion, leaving the system empty (see Heyman (1977)). Another policy is D-policy, according to which the service facility reopens as soon as the total workload (after each service completion leaving the system empty) exceeds a critical level D (see Sivazlian (1979)). (Refer to Tijms (1986) for analyses of control models under these policies by renewal reward processes.)

Studies relate to determination of optimal policies according to specified objectives—for example, minimization of average waiting time and minimization of average cost as per specific cost structure.

As did Heyman (1968) and Bell (1971), Sobel (1969) also considers cost structure along with start-up and shut-down costs for a single-server system. McGill considers a general switching-cost model for a system with a variable number of exponential servers. Bell (1975) considers the average cost criterion, while discounted cost over a finite horizon for a c-server Markovian model is considered by Huang *et al.* (1977). Assuming that the cost structure includes customer-holding (waiting) cost and service-channel-holding (server running)

cost as well as linear-switching cost, Szarkowicz and Knowles (1985) consider optimal control of an $M/M/c$ system and show that control-limit policy is optimal under less restrictive conditions. They also use the dynamic programming formulation.

Remark: Control problems for bulk-service queues have been considered, among others, by Kosten (1967), Deb and Serfozo (1973), Ignall and Kolesar (1974), Weiss (1979), Powell and Humblet (1986), Krishnamoorthy and Ushakumari (2000), and so on. Consider a Poisson input system operated under the following policy. When the server is available and there are fewer than Q customers waiting, the service does not begin. If there are Q or more waiting and the server is available, service begins with all those waiting. This is an infinite-capacity bulk-service queue of the type $M/G(Q,\infty)/1$. Assume that a fixed start-up cost K is incurred each time service is initiated, and the waiting cost per customer is h per unit of time. Deb and Serfozo (1973) show that the optimal type of policy is the control-limit policy, which requires that service begins if and only if the number of waiting customers is as large as Q. Weiss (1979) finds expressions for average long-run cost per unit time and for the optimal Q. He applies renewal theory on the assumption that the epoch at which service begins each time (with none left in the queue because of infinite server capacity) are regeneration points. The situation will not be the same for a general-bulk-service queue with *finite* server capacity.

Since mass-transit vehicles are sort of natural batch servers, and since a shuttle between two destinations can be considered as a single-server system, this kind of bulk-service model can be used as a model for such systems. Similarly, multiserver models could also fit as realistic models for mass-transportation systems. In view of this, control problems for batch-service systems have assumed importance.

For a brief survey, see, for example, Medhi (1984a,b). Powell and Humblet (1986) consider unified treatment for a number of vehicle-dispatch strategies, such as vehicle-holding strategy (general bulk-service rule), vehicle-cancellation strategy, as well as a combination of these two strategies. A Markov chain approach is considered. They define Q_n as the steady-state queue length at the nth dispatch instant (instant of nth service completion) and show that, under certain conditions, the Markov chain $\{Q_n, n \geq 0\}$ is ergodic, and they obtain the PGF of the steady-state queue length at dispatch instants in terms of the PGFs of associated variables arising out of the control strategy. They also obtain a decomposition type of result

$$Q(z) = Q_1(z)\,B(z),$$

where $Q(z)$ and $Q_1(z)$ are the PGFs of the queue length at dispatch instants of corresponding queues with and without control strategy, respectively, and $B(z)$ is a function connected with the control strategy.

Nobel (1998) studies hysteretic (with two levels of service changing from one service to the other and vice versa) and heuristic control of queues with batch Poisson arrivals. See also Nobel and Tijms (1999).

Since computer communication networks are modeled as queueing networks, the question of optimal flow control in queueing networks has been receiving attention. (See, for example, Lazar (1983) and Sauer and Chandy (1981).)

All these studies show the importance and use of control and design of queues in practical applications.

8.5 Retrial Queueing System

It may often happen that a telephone caller when dialing a number (say, public utility) gets a busy signal (finds the telephone facility busy because it is already engaged by another caller). In such cases, the caller repeats his call after a random amount of time; other callers also do the same. These callers become *sources* of repeated calls and remain in what is termed in an *Orbit*, while a fresh caller (called a *primary caller*) who finds the facility free immediately gets the facility. These considerations bring into focus the need for a new queueing system, which is called the *retrial queueing system*. The theory of retrial queues is now considered an important part of queueing theory as well as teletraffic theory. The literature on retrial queues, mainly confined to sections in books on queueing theory and research papers in journals, has been growing. There is a significant contribution to the literature from Russian researchers, mostly in Russian.

Specific results appeared in the 1950s, and survey papers were published by Yang and Templeton (1987), Falin (1990), Kulkarni and Liang (1997), and so on. A monograph by Falin and Templeton (1997) appeared recently. A classified bibliography (Artalejo, 1999b) focuses on progress during 1990–1999 and supplements his earlier bibliography (1999a) of publications prior to 1990. These papers along with the monograph constitute a very exhaustive source of references on this topic. A very large number of researchers have contributed to the development of the subject area.

Here we consider a few basic models in order to give an idea of the structure of retrial queueing systems. For more details, readers may look into the above-mentioned literature and also to the references cited therein (especially in Falin and Templeton (1997)).

8.5.1 Retrial queues: model description

For simplicity, let us first confine ourselves to a single-server system in steady state. Consider a queueing system where calls (customers) from outside arrive in a Poisson stream with rate λ and demand service from a facility with a service

time S having DF $B(.)$ and rate μ. These calls are called *primary calls.* If the server is free, the arriving customer (primary call) begins service immediately. If the server is not free, the arriving customer (the arriving call) enters an orbit and attempts to get service after a random amount of time, called *retrial time.* If, at the end of the retrial time, he finds the server free, he is admitted into service and leaves the system after service. Every arriving call from the orbit that does not find the server free returns to the orbit, remains in the orbit (in a sort of queue), and becomes a source of repeated calls. Each source, independent of other sources, produces a Poisson stream of repeated calls with rate θ. A repeated call after the nth unsuccessful attempt to get the server free may return to the orbit with probability α_n or may not return to the orbit and may leave the system with probability $(1 - \alpha_n, n \geq 1)$. If $\alpha_n = 1$, the source is called persistent, and if $\alpha_n = \alpha < 1$, the source is said to be nonpersistent or impatient.

We shall number the calls in the order of service. Suppose that the $(i-1)$th call completes its service at instant τ_{i-1} and the server becomes free. A primary call or a source from the orbit may be the next call to be taken for service. If there are n customers in the orbit (of repeated calls), the probability that the next call (ith) is a primary call has probability $\frac{\lambda}{\lambda+n\theta}$, and the probability that the next call is a repeated call from the orbit has probability $\frac{n\theta}{\lambda+n\theta}$. Even with the free server at the instant τ_{i-1}, when the server becomes free, a call from the source in orbit (not being familiar with the state of the service facility) may not join the facility immediately but may join after some interval R_i, so the instant of completion of service of the ith call (if no primary call arrives in the interval R_i) will be $\tau_i + R_i + S_i$, S_i being the service time of the ith call, and the server becomes free again. Assume that the arrival stream of primary calls, the service times, and the retrial times are mutually independent.

It is also assumed that the customers are persistent, that there is no space in the service facility except for the one in service, and that the orbit has infinite space. The systems, where service facility has additional space or the orbit has limited waiting space, have also been considered in the literature.

At time t, let $N(t)$ be the number of sources of repeated calls and $C(t)$ be the number of busy servers. We consider the bivariate process $\{C(t), N(t)\}, t \geq 0$, where $C(t) = 1$ or 0 according as the server is busy or idle; the process will be called the CN process. If service time S with DF $B(.)$ is exponential, then $\{C(t), N(t)\}$ is Markovian. In case $B(.)$ is not exponential, and $C(t) = 1$, we consider supplementary variable $\xi(t)$ being the elapsed service time of the call in service at time t.

The stability (existence of steady state) of the CN process for different models has been investigated, and conditions for the same have been obtained. Here we examine the steady-state distributions of the CN process assuming that these exist for the models considered. For the $M/M/1$ retrial queue, stationary distribution exists, if $\rho = \lambda E(S) = \lambda/\mu < 1$ and we assume that it holds.

Let C and N be the number of customers in the service facility and in the orbit, respectively, in steady state. Denote $p_{in} = P\{C = i, N = n\}, i = 0, 1, n \geq 0$.

8.5.2 Single-server model: $M/M/1$ retrial queue

Here the arrival stream is Poisson (with rate λ), service time is exponential (with rate μ), and retrial time is also exponential (with rate θ).

Theorem 8.10. *For an $M/M/1$ retrial queue, p_{in} are given by*

$$p_{0,n} = \frac{\rho^n}{n!\theta^n} p_{0,0} \prod_{i=0}^{n-1} (\lambda + i\theta), \quad n \geq 1 \tag{8.5.1}$$

$$\text{and} \quad p_{1,n} = \frac{\rho^{n+1}}{n!\theta^n} p_{0,0} \prod_{i=1}^{n} (\lambda + i\theta), \quad n \geq 0, \tag{8.5.2}$$

where $p_{0,0} = (1-\rho)^{\lambda/\theta+1}$.

Proof: Here CN is a Markov process.
From state $(0, n)$ transitions only to the following states are possible.

(i) state $(1, n)$ with rate $\lambda p_{0,n}$ (with the arrival of a primary call), and

(ii) state $(1, n-1)$ with rate $n\theta p_{0,n}$ (with the commencement of service of one of the n sources in the orbit, as the server is free).

Again state $(0, n)$ can be reached with transitions only from the following states: state $(1, n)$ with completion of service of the call in service with rate $\mu p_{1,n}$.

Since, for equilibrium, rate up = rate down (rate out = rate in), we get

$$(\lambda + n\theta)\, p_{0,n} = \mu p_{1,n}. \tag{8.5.3}$$

From state $(1, n)$ transitions only to the following states are possible.

(i) state $(1, n+1)$ with rate $\lambda p_{1,n}$ (with arrival of a primary call which goes to the orbit as he finds the server busy), and

(ii) state $(0, n)$ with rate $\mu p_{1,n}$ (with the service completion of the call in service).

Again state $(1, n)$ can be reached with transitions only from the following states.

(i) state $(1, n-1)$ with rate $\lambda p_{1,n-1}$ (with the arrival of a primary call which is admitted to service as the server is free).

(ii) state $(0, n)$ with rate $\lambda p_{0,n}$ (with the arrival of a primary call which is admitted to service as the server is free).

(iii) state $(0, n+1)$ with rate $(n+1)\theta p_{0,n+1}$ (with the commencement of service of one of the $(n+1)$ sources in the orbit).

Thus, we have

$$(n+\mu)\,p_{1,n} = \lambda p_{1,n-1} + \lambda p_{0,n} + (n+1)\theta p_{0,n+1}. \tag{8.5.4}$$

Solution of (8.5.3) and (8.5.4), together with the normalization condition,

$$\sum_{n=0}^{\infty} p_{0,n} + \sum_{n=0}^{\infty} p_{1,n} = 1 \tag{8.5.5}$$

will give $p_{0,n}$ and $p_{1,n}$. Let

$$P_0(z) = \sum_{n=0}^{\infty} p_{0,n} z^n \quad \text{and} \quad P_1(z) = \sum_{n=0}^{\infty} p_{1,n} z^n$$

be the partial generating functions of $\{p_{0,n}\}$ and $\{p_{1,n}\}$, respectively. Multiplying (8.5.3) and (8.5.4) by different powers of z and adding, we get

$$\lambda P_0(z) + \theta z P_0'(z) = \mu P_1(z) \tag{8.5.6}$$

$$(\mu + \lambda - \lambda z) P_1(z) = \lambda P_0(z) + \theta P_0'(z). \tag{8.5.7}$$

Eliminating $P_1(z)$ from the above two relations, we get

$$P_0'(z) = \frac{\lambda\rho}{\theta(1-\rho z)} P_0(z)$$

and integrating

$$P_0(z) = \frac{C}{(1-\rho z)^{\lambda/\theta}},$$

where C is a constant. Substituting in (8.5.6), one gets

$$\begin{aligned} P_1(z) &= \rho P_0(z) + \frac{\rho^2 z}{1-\rho z} P_0(z) \\ &= \frac{C\rho}{(1-\rho z)^{\frac{\lambda}{\theta}+1}}. \end{aligned} \tag{8.5.8}$$

The normalization condition (8.5.5) becomes

$$P_0(1) + P_1(1) = 1,$$

which gives

$$C = (1-\rho)^{\frac{\lambda}{\theta}+1}. \tag{8.5.9}$$

Now comparing the coefficients of z^n in $P_i(z)$ from both sides, one gets $p_{i,n}$ for $i = 0, 1$ and $n = 0, 1, 2, \ldots$ ■

Note 1: The PGF of the number Q of sources of repeated calls in orbit is given by

$$Q(z) = \sum_{n=0}^{\infty} P\{Q = n\}z^n = P_0(z) + P_1(z)$$

$$= (1 + \rho - \rho z)\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}+1}. \tag{8.5.10}$$

$P\{Q = n\}$ can be found by expanding the RHS in powers of z and comparing powers of z from both sides.

We get

$$E\{Q\} = Q'(1) = \frac{\rho(\lambda + \rho\theta)}{(1-\rho)\theta} \tag{8.5.11a}$$

and

$$\text{var}\{Q\} = Q''(1) - [Q'(1)]^2$$

$$= \frac{\rho(\lambda + \rho\theta + \rho^2\theta - \rho^3\theta)}{(1-\rho)^2\theta}. \tag{8.5.11.b}$$

By Little's theorem, the expected waiting (queueing) time is

$$E(W_Q) = \frac{\rho}{1-\rho}\left(\frac{1}{\theta} + \frac{1}{\mu}\right)$$

and the expected response time is

$$E(W) = E(W_Q) + \frac{1}{\mu} = \frac{1}{1-\rho}\left(\frac{1}{\mu} + \frac{\rho}{\theta}\right)$$

Note 2: The blocking probability p_1 is the probability that the server is busy; it is given by

$$p_1 = \sum_{n=1}^{\infty} p_{1,n} = P_1(1) = \rho$$

and the mean number of busy servers is ρ.

These should be intuitively clear.

Note 3: The PGF of the number N of customers in the system (in the orbit and in service, if any) is given by

$$P(z) = \sum_{n=0}^{\infty} P\{N = n\}z^n = P_0(z) + zP_1(z) = \left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\theta}+1}, \tag{8.5.12}$$

whence $P\{N = n\}$ can be found by expanding the RHS in powers of z and comparing coefficients of z. We get

$$E\{N\} = P'(1) = \frac{\rho(\lambda + \theta)}{(1 - \rho)\theta} \tag{8.5.13}$$

$$\text{and} \quad \text{var}\{N\} = \frac{\rho(\lambda + \theta)}{(1 - \rho)^2\theta}.$$

Note 4: Taking $\theta \to \infty$, we get the standard $M/M/1$ queue with $p_{0,n} = 0$, $n \neq 0$,

$$p_{1,n} = P\{N = n + 1\} = (1 - \rho)\rho^{n+1}, \quad \text{and} \quad p_{0,0} = p_0 = (1 - \rho).$$

Note 5: For $\theta = 0$, the system behaves as a loss system $M/M/1/1$.

For results on stability conditions, waiting-time process, departure process, busy period, transient-state distribution, and other related topics, reference may be made to Falin and Templeton (1997) and references therein.

8.5.3 $M/G/1$ retrial queue

Consider a single-server retrial queue with general service time S, having DF $B(.)$ mean $\frac{1}{\mu}$ and LST $B^*(.)$. Consider that the system is in steady state and that there is no space before the server except for the one in service, if any. Here we introduce a supplementary variable—the elapsed service time $\xi(t)$ at time t of the customer in service, when the server is busy.

The PGF of the number of primary calls that arrive during the service time of a call is given by $K(z) = B^*(\lambda - \lambda z)$. The hazard rate function $r(x)$ of S is given by

$$r(x) = B'(x)/\{1 - B(x)\}.$$

Denote

$$p_{0,n}(t) = P\{C(t) = 0, N(t) = n\} \tag{8.5.14a}$$

$$p_{1,n}(t, x) = P\{C(t) = 1, N(t) = n, \xi(t) < x\}. \tag{8.5.14b}$$

As $t \to \infty$, let

$$p_{0,n}(t) \equiv p_{0,n}, \quad p_{1,n}(t, x) \equiv p_{1,n}(x), \quad p_{1,n} \equiv \int_0^\infty p_{1,n}(x)dx. \tag{8.5.14c}$$

Theorem 8.11. *For an $M/G/1$ Retrial Queue*

$$P_0(z) \equiv \sum_{n=0}^{\infty} p_{0,n} z^n$$

$$= (1 - \rho) \exp\left\{\frac{\lambda}{\theta} \int_1^z \frac{1 - K(u)}{K(u) - u} du\right\} \tag{8.5.15}$$

and

$$P_1(z,x) \equiv \sum p_{1,n}(x)z^n$$
$$= \lambda\left[\frac{1-z}{K(z)-z}\right]P_0(z)[1-B(x)]e^{-\lambda(1-z)x}. \quad (8.5.16)$$

Further

$$P_1(z) = \sum_{n=0}^{\infty} p_{1,n}z^n = \frac{1-K(z)}{K(z)-z}\,P_0(z). \quad (8.5.17)$$

Proof: Proceeding as in the case of the standard $M/G/1$ queue (Section 6.3.3), we get the following state equations (corresponding to the equations (6.3.23)–(6.3.24))

$$(\lambda + n\theta)\,p_{0,n} = \int_0^{\infty} p_{1,n}(x)r(x)dx \quad (8.5.18)$$

$$p'_{1,n}(x) = -(\lambda + r(x))\,p_{1,n}(x) + \lambda p_{1,n-1}(x). \quad (8.5.19)$$

From the boundary condition, one gets

$$p_{1,n} = \lambda p_{0,n} + (n+1)\theta p_{0,n+1}. \quad (8.5.20)$$

Multiplying by different powers of z and adding, we get, in terms of the generating functions

$$\lambda P_0(z) + \theta z P_0'(z) = \int_0^{\infty} P_1(z,x)r(x)dx \quad (8.5.21)$$

$$\frac{\partial}{\partial x}P_1(z,x) = -(\lambda - \lambda z + r(x))\,P_1(z,x) \quad (8.5.22)$$

$$P_1(z,0) = \lambda P_0(z) + \theta P_0'(z). \quad (8.5.23)$$

Solving (8.5.22), we get

$$P_1(z,x) = P_1(z,0)[1-B(x)]e^{-(\lambda-\lambda z)x}. \quad (8.5.24)$$

Substituting in (8.5.21), we get

$$\lambda P_0(z) + \theta z P_0'(z) = P_1(z,0)\int_0^{\infty}[1-B(x)]e^{-(\lambda-\lambda z)x}r(x)dx$$
$$= P_1(z,0)\int_0^{\infty} e^{-(\lambda-\lambda z)x}\,dB(x)$$
$$= P_1(z,0)\,B^*(\lambda-\lambda z) = P_1(z,0)\,K(z). \quad (8.5.25)$$

Eliminating $P_0'(z)$ from (8.5.23) and (8.5.25), we get

$$\lambda P_0(z) + z[P_1(z,0) - \lambda P_0(z)] = P_1(z,0)K(z)$$

or

$$P_1(z,0) = \frac{\lambda(1-z)}{K(z)-z} P_0(z). \tag{8.5.26}$$

Substituting in the RHS of (8.5.24), we get (8.5.16).

Eliminating $P_1(z,0)$ from (8.5.25) and (8.5.26), we get

$$\lambda P_0(z) + \theta z P_0'(z) = \frac{\lambda(1-z)}{K(z)-z} K(z) P_0(z)$$

and, on simplification,

$$\frac{P_0'(z)}{P_0(z)} = \frac{\lambda(1-K(z))}{\theta(K(z)-z)}.$$

Solving, and using $P_0(1) = 1 - \rho$, we get (8.5.15).

Integrating (8.5.16) WRT x from 0 to ∞, we get

$$\begin{aligned}
P_1(z) &= \sum_{n=0}^{\infty} p_{1,n} z^n \\
&= \sum_{n=0}^{\infty} \left\{ \int_0^{\infty} p_{1,n}(x)\,dx \right\} z^n \\
&= \int_0^{\infty} P_1(z,x)\,dx \\
&\quad \text{(assuming the validity of change of order of summation and integration)} \\
&= \frac{\lambda(1-z)}{K(z)-z} P_0(z) \int_0^{\infty} e^{-\lambda(1-z)x}\{1 - B(x)\}\,dx \\
&= \frac{\lambda(1-z)}{K(z)-z} P_0(z) \cdot \frac{1 - B^*(\lambda - \lambda z)}{\lambda(1-z)} \\
&= \frac{1-K(z)}{K(z)-z} P_0(z),
\end{aligned}$$

which is (8.5.17). ■

Notes:

(1) The PGF of the number Q of sources of repeated calls is given by

$$\begin{aligned} Q(z) &\equiv \sum_{k=0}^{\infty} P(Q = k) z^k \\ &= P_0(z) + P_1(z) \end{aligned}$$

with mean

$$E(Q) \equiv Q'(1) = \frac{\lambda^2}{1-\rho}\left[\frac{E(S)}{\theta} + \frac{E(S^2)}{2}\right].$$

By Little's theorem, the mean waiting time is

$$E(W) = \frac{\lambda}{1-\rho}\left[\frac{E(S)}{\theta} + \frac{E(S^2)}{2}\right].$$

(2) The probability that the server is in state 0 (idle), 1 (busy) is given by

$$p_0 = 1 - \rho, \qquad p_1 = \rho.$$

(3) The PGF of the number N of customers in the system is given by

$$\begin{aligned} N(z) &\equiv \sum_{n=0}^{\infty} P\{N = n\} z^n \\ &= P_0(z) + z P_1(z). \end{aligned}$$

We have $E\{N\} = E\{Q\} + \rho$.

(4) As the retrial rate $\theta \to \infty$, the retrial system converges to the standard $M/G/1$ system. For $\theta = 0$, the system is equivalent to the Erlang loss system $M/G/1/1$.

(5) The interdeparture interval τ_i consists of two parts: the idle period R_i (after departure of the ith customer to the commencement of service of the next customer) and S_{i+1}, the service time of the $(i+1)$th customer. In steady state the arrival rate must be equal to the departure rate. So $E\{\tau_i\} = 1/\lambda$. Again

$$E\{\tau_i\} = E\{R_i\} + E\{S_{i+1}\} = E\{R_i\} + 1/\mu,$$

whence $E\{R_i\} = (1-\rho)/\lambda$.

(6) Note that the expected idle period is independent of the rate θ of the repeated calls and the number in the orbit.

Let N_{i-1} be the number of customers in the orbit at the instant that the $(i+1)$th call completes its service. R_i depends on N_{i-1}. R_i has the conditional distribution

$$P\{x \le R_i < x + dx \mid N_{i-1} = n\} = (\lambda + n\theta) e^{-(\lambda + n\theta)x}\, dx$$

having mean $E\{R_i \mid N_{i-1} = n\} = 1/(\lambda + n\theta)$.

While the mean server idle time is $(1-\rho)/\lambda$, the LST $I^*(s)$, the idle time of the $M/M/1$ retrial queue is given by

$$I^*(s) = E\{e^{-sI}\} = \frac{1}{\lambda+s}\left[\lambda + s\int_0^1 x^{(s+\lambda)/\theta}\, dQ(x)\right],$$

where $Q(x)$ is the MGF of the limiting distribution of the number of customers at an arbitrary departure epoch (Yang and Templeton, 1987).

8.5.4 Multiserver models

8.5.4.1 M/M/c retrial queue

Here we have a group of c parallel servers and the calls (primary calls/customers) arrive in a Poisson stream with rate λ. Such a call is immediately admitted for service, if any of the c-servers is free, and departs after service completion. If it does not find any free server, it goes to the orbit and becomes a source of repeated calls to retry after a random amount of time—the retrial times (interval between successive retrials) are exponential with rate θ. The service times are assumed to be exponential with DF $B(.)$ with mean $1/\mu$, which for simplicity will be assumed to be 1 (so that $\rho = \lambda$). It is assumed that the interarrival times (of primary calls), service times, and retrial times are mutually independent. Consider that steady states exist.

The system can be described by the bivariate process $\{C, N\}$, where C is the number of busy servers and N is the number of customers (sources) in the orbit. The process is Markovian. Here the number of busy servers C can take values $0, 1, \ldots, c$, and the number of sources in the orbit N can take values $0, 1, \ldots$. The transition rates $q_{(ij)(mn)}$ from state (i, j) to state (m, n) in an infinitesimal interval are as follows.

For $0 \le i \le c-1$

$$\begin{aligned} q_{(i,j)(mn)} &= \lambda, \quad \text{for } (m,n) = (i+1, j) \\ &\quad \text{(arrival of a primary call)} \\ &= i, \quad \text{for } (m,n) = (i-1, j) \\ &\quad \text{(completion of service of one of the } i \text{ customers in service)} \\ &= j\theta, \quad \text{for } (m,n) = (i+1, j-1) \\ &\quad \text{(one of the } j \text{ sources in the orbit goes to one of the free servers} \\ &\quad \text{available, which increases the number of busy servers by 1)} \\ &= -(\lambda + i + j\theta), \quad \text{for } (m,n) = (i,j) \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{8.5.27}$$

For $i = c$,

$$\begin{aligned} q_{(c,j)(m,n)} &= \lambda, \quad \text{for } (m,n) = (c, j+1) \\ &= c, \quad \text{for } (m,n) = (c-1, j) \\ &= -(\lambda + c), \quad \text{for } (m,n) = (c, j) \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{8.5.28}$$

8.5.4.2 The case $c = 2(\mu = 1)$

Denote

$$p_{i,j} = P\{c = i, N = j\}, \quad \begin{matrix} i = 0, 1, 2 \\ j = 0, 1, 2, \ldots \end{matrix}$$

These probabilities are obtained as solutions of the Kolmogorov equations

$$(\lambda + j\theta)\, p_{0j} = p_{1j} \tag{8.5.29}$$

$$(\lambda + 1 + j\theta)\, p_{1j} = \lambda p_{0j} + (j+1)\theta p_{0,j+1} + 2p_{2,j} \tag{8.5.30}$$

$$(\lambda + 2)\, p_{2j} = \lambda p_{1j} + (j+1)\theta p_{1,j+1} + \lambda p_{2,j-1} \tag{8.5.31}$$

with the normalization condition

$$\sum_{j=0}^{\infty} (p_{0j} + p_{1j} + p_{2j}) = 1. \tag{8.5.32}$$

Eliminating p_{ij} from (8.5.29) and (8.5.30), one gets

$$2p_{2,j} = [(\lambda + j\theta)^2 + j\theta]\, p_{0,j} - (j+1)\theta p_{0,j+1}, \tag{8.5.33}$$

which gives $p_{2,j}$ in terms of $p_{0,j}$ and $p_{0,j+1}$.

Using expressions for $p_{2,j}$ and $p_{2,j-1}$ in (8.5.31), we get a relationship between $p_{0,j}$, $p_{0,j+1}$ and $p_{0,j-1}$ which can be written as the following recursive relation

$$\begin{aligned} &\lambda[(\lambda + j\theta)^2 + j\theta]\, p_{0,j} - (j+1)\theta[2 + 3\lambda + 2(j+1)\theta]\, p_{0,j+1} \\ &\quad = \lambda[\{\lambda + (j-1)\theta\}^2 + (j-1)\theta]\, p_{0,j-1} - j\theta[2 + 3\lambda + 2j\theta]\, p_{0,j}. \end{aligned}$$

From this, we have, for all $j(\geq 1)$

$$\lambda[\{\lambda + (j-1)\theta\}^2 + (j-1)\theta]\, p_{0,j-1} - j\theta[2 + 3\lambda + 2j\theta]\, p_{0,j} = 0. \tag{8.5.34}$$

Thus, for all $j(\geq 1)$

$$p_{0,j} = \frac{\lambda}{j\theta} \cdot \frac{[\{\lambda + (j-1)\theta\}^2 + (j-1)\theta]}{2 + 3\lambda + 2j\theta}\, p_{0,j-1}. \tag{8.5.35}$$

Writing $j-1, j-2, \ldots 2, 1$ for $j-1$ in the RHS, we get

$$p_{0,j} = \frac{\lambda^j}{j!\theta^j}\left\{\prod_{k=0}^{j-1}\frac{(\lambda+k\theta)^2+k\theta}{2+3\lambda+2(k+1)\theta}\right\}p_{0,0}. \tag{8.5.36}$$

From (8.5.29) and (8.5.33) we get

$$p_{1,j} = (\lambda+j\theta)\frac{\lambda^j}{j!\,\theta^j}\left\{\prod_{k=0}^{j-1}\frac{(\lambda+k\theta)^2+k\theta}{2+3\lambda+2(k+1)\theta}\right\}p_{0,0} \tag{8.5.37}$$

$$p_{2,j} = \{\lambda+1+(j+1)\theta\}\frac{\lambda^j}{j!\theta^j}\left\{\prod_{k=0}^{j}\frac{(\lambda+k\theta)^2+k\theta}{2+3\lambda+2(k+1)\theta}\right\}p_{0,0}. \tag{8.5.38}$$

Using (8.5.32), $p_{0,0}$ can be obtained as

$$[p_{0,0}]^{-1} = \sum_{j=0}^{\infty}\frac{\lambda^j}{j!\theta^j}\left\{\prod_{k=0}^{j-1}\frac{(\lambda+k\theta)^2+k\theta}{2+3\lambda+2(k+1)\theta}\right\}$$
$$\times\left[\lambda+1+j\theta+\frac{\{\lambda+1+(j+1)\theta\}\{(\lambda+j\theta)^2+j\theta\}}{2+3\lambda+2(j+1)\theta}\right]. \tag{8.5.39}$$

The generating functions $P_i(z) = \sum_{j=0}^{\infty} p_{ij}z^j$, $i = 0, 1, 2$ and the probability $p_{0,0}$ can be expressed in terms of hypergeometric functions defined as

$$F(a, b, c : x) = \sum_{j=0}^{\infty}\frac{x^j}{j!}\prod_{k=0}^{j-1}\frac{(a+k)(b+k)}{(c+k)}.$$

While the distribution has been found by Jonin and Sedol (1970), representation of the generating function in terms of hypergeometric functions is by Hanschke (1987).

A recursive algorithm to compute the distribution has been proposed by Keilson *et al.* (1968).

8.5.4.3 General case: $M/M/c$ retrial queue ($\mu = 1$)

Let $p_{ij} = P\{C = i, N = j\}$, $i = 0, 1, 2, \ldots, c$; $j = 0, 1, 2, \ldots$. The CN process is Markovian. These probabilities satisfy the Kolmogorov equations

$$(\lambda+i+j\theta)\,p_{ij} = \lambda p_{i-1,j} + (j+1)\theta p_{i-1,j+1} + (i+1)\,p_{i+1,j}, \quad 0 \le i \le c-1$$
$$(\lambda+c)\,p_{cj} = \lambda p_{c-1,j} + (j+1)\theta p_{c-1,j+1} + \lambda p_{c,j-1}$$

and the normalizing condition $\sum_{j=0}^{\infty}(p_{0j} + p_{1j} + \cdots + p_{cj}) = 1$.

The solution of the above for $c > 2$ is too complicated, and no explicit formulas have been found for the probabilities. Using generating functions, we can obtain some of the performance measures. Some approximate methods have been suggested. Neuts and Rao (1990) consider numerical solutions using the matrix geometric approach.

8.5.5 Model with finite orbit size

Here the size of the orbit (which accommodates the sources of repeated calls) is finite—say, K—that is, the orbit can hold only K sources. The corresponding model is also called the truncated model. Assume that the arrival process is Poisson, retrial times are exponential (with mean $\frac{1}{\theta}$), and the service times (of the parallel servers) are exponential (with mean 1) and that the three sets are mutually independent. Then the process is Markovian.

The CN process being finite, it is always ergodic (steady state exists). For simplicity we drop the letter K and use the same notation, as in the general case with infinite orbit size. The transition probabilities will be the same as the main model with c servers except for the boundary state $i = c$ and $j = K$. These are

(a) for $1 \leq i \leq c-1, 0 \leq j \leq K$

$$q_{(ij)(mn)} = \begin{cases} \lambda, & \text{for } (m,n) = (i+1, j) \\ i, & \text{for } (m,n) = (i-1, j) \\ j\theta, & \text{for } (m,n) = (i+1, j-1) \\ -(\lambda + i + j\theta), & \text{for } (m,n) = (i,j) \\ 0, & \text{otherwise} \end{cases}$$

(b) for $i = c, 0 \leq j \leq K-1$

$$q_{(cj)(mn)} = \begin{cases} \lambda, & \text{for } (m,n) = (c, j+1) \\ c, & \text{for } (m,n) = (c-1, j) \\ -(\lambda + c), & \text{for } (m,n) = (c,j) \\ 0, & \text{otherwise} \end{cases}$$

(c) for $i = c, j = K$

$$q_{(cK)(mn)} = \begin{cases} c, & \text{for } (m,n) = (c-1, K) \\ -c, & \text{for } (m,n) = (c,K) \\ 0, & \text{otherwise} \end{cases}$$

The steady-state probabilities $p_{ij} = P\{C = i, N = j\}, 0 \le i \le c, 0 \le j \le K$ are given as the solutions of the following Kolmogorov equations:

$$\begin{aligned}
(\lambda + i + j\theta)\, p_{ij} &= \lambda p_{i-1,j} + (j+1)\theta p_{i-1,j+1} + (i+1)\, p_{i+1,j}, \\
&\qquad 0 \le i \le c-1, \;\; 0 \le j \le K-1 \\
(\lambda + i + K\theta)\, p_{iK} &= \lambda p_{i-1,K} + (i+1)\, p_{i+1,K}, \quad 0 \le i \le c-1 \\
(\lambda + c)\, p_{cj} &= \lambda p_{c-1,j} + (j+1)\theta p_{c-1,j+1} + \lambda p_{c,j-1}, \quad 0 \le j \le K-1 \\
c p_{cK} &= \lambda p_{c-1,K} + \lambda p_{c,K-1}
\end{aligned}$$

together with the normalizing condition

$$\sum_{i=0}^{c} \sum_{j=0}^{K} p_{ij} = 1.$$

Let

$$P_i(z) = \sum_{j=0}^{K} p_{ij} z^j, \quad 0 \le i \le c$$

and

$$P(x, z) = \sum_{i=0}^{c} P_i(z) x^i.$$

The above equations can then be conveniently expressed in terms of the above generating functions. From these relations, performance measures can be obtained.

The state equations are finite in number. Algorithms for numerical computation of the solutions can be worked out. A recursive algorithm for calculation of $p_{i,j}$ has also been put forward (see Falin and Templeton (1997)).

8.5.6 Other retrial queue models

These are several modifications of these models of retrial queues. In a single-server model with batch arrival (with compound Poisson input), if the arrivals find the server free, one of the batch goes to the service facility, whereas the others of the batch go to the orbit. Some of the other models considered are models with impatient customers, with multiclass customers, with finite buffer (the service facility having some extra space), and with server breakdown and repairs, server vacation, and so on.

Advanced multiserver models are models with impatient customers, priority customers, a finite input source of primary calls, and so on. Estimation of retrial rates and other estimation problems are also considered.

Readers are referred to the monograph by Falin and Templeton (1997) and to the references therein and also to Artalejo's (1999b) classified bibliography, which gives classification according to two criteria—the first one is an author index, and the second one is an exhaustive subject classification, to which interested readers may refer.

There are several open problems in this topic (Kulkarni and Liang (1997) mention some of them). One of the most important open problems concerns the process and the performance measures thereof in the case of general retrial intervals. Other problems arise in connection with stability of the system, busy period analysis, retrial queueing networks, transient analysis, and so on. Retrial queueing systems are important and interesting ("but not easy," as observed by Templeton: Retrial Queues in *TOP* 7 # 2, 1999, p. 351).

8.6 Emergence of a New Trend in Teletraffic Theory

8.6.1 Introduction

Queueing models have been very successfully applied in classical teletraffic theory. These have led to accurate analysis of modern communication and computer systems. Queueing theory has played a very significant role in modeling voice calls. The validity of Poisson processes to describe voice call arrivals in a circuit-switched voice network has been indicated by actual measurement studies of such voice calls. Another reason for the applicability of queueing models is the robustness property of many important queueing results—for example, the insensitivity of the Erlang loss formula to service time distribution. Voice traffic is homogeneous and has limited variability.

With the advent of faxes (in the 1980s) and the Internet (during the last decade), which also use the same telephone networks (POTS: Plane Old Telephone System), the nature of traffic has changed dramatically. As a result, packet-switched networks have gained importance over circuit-switched networks. Data traffic (where modems do the talking instead of humans) are very much different from voice traffic in their statistical characteristics. These are more variable and vary from extremely short to extremely long and from an extremely low rate to an extremely high rate. While ordinary voice calls are of average duration of only a few minutes, data traffic (comprising faxes and the Internet) can run for several hours together. Data traffics do not come in a steady rate but come with starts and fits, with lulls in between: They are bursty. In contrast to voice traffic, which is homogeneous, burstiness, involving a wide range of length or time scales, is experienced with modern-day data traffic.

Traffic data as encountered now indicate different kinds of statistical characteristics. These are (1) distributions of traffic processes decay more slowly than exponential, and (2) autocorrelation functions are found to exhibit hyperbolic decay rather than exponential decay, indicating long-range dependence

rather than short-range dependence. Variabilities occur in both space and time. Spatial variability of the traffic processes is captured by long tail (heavy tail) distributions and temporal variability by long-range dependence. These are termed (by Mandelbrot (1965)) the *Noah Effect* and *Joseph Effect*, respectively.

We shall discuss here one of the characteristic feature: the infinite variance syndrome, manifested by long tail (heavy-tail) distributions.

In recent years, researchers have been engaged in the study of these types of traffic processes in the context of queueing analysis. There is a resurgence of interest in the use of distributions with " heavy tails" as inputs to queueing models. An issue of *Queueing Systems* (Vol. **33**, 1999) was entirely devoted to "Queues with heavy-tailed distributions."

A brief survey is given in Medhi (2001).

8.6.2 Heavy-tail distributions

Consider exponential distribution having DF $F(x) = 1 - e^{-\lambda x}$ and CDF (complementary distribution function) $F_c(x) = 1 - F(x)$. We have

$$\lim_{x\to\infty} \frac{F_c(x+y)}{F_c(x)} = \lim_{x\to\infty} e^{-\lambda y} = e^{-\lambda y}, \quad x \geq 0, \quad y \geq 0,$$

which is dependent on y. The distribution is said to have a short tail (light tail or lean tail). Consider the Pareto distribution having CDF

$$F_c(x) = \left(\frac{a}{x}\right)^r, \quad a > 0, \quad r > 0; \quad x \geq a.$$

We have

$$\lim_{x\to\infty} \frac{F_c(x+y)}{F_c(x)} \longrightarrow 1, \quad \text{for all } y \geq 0. \tag{8.6.1}$$

This may also be denoted as $F_c(x + y) \sim F_c(x)$, where $a(x) \sim b(x)$ imply that

$$a(x)/b(x) \to 1 \quad \text{as } x \to \infty.$$

This distribution is long tailed (or heavy tailed or fat tailed). Its intuitive meaning is that if the RV X ever exceeds a large value, then it is likely to exceed any larger value as well.

The nth central moment of the Pareto distribution is

$$\mu_n' = \frac{ra^n}{r-n}, \quad 1 \leq n < r.$$

The expected value is $ra/(r-1)(if\ r > 1)$, and the variance is

$$\frac{ra^2}{(r-1)^2(r-2)}, \quad r > 2.$$

For $r > 1$, the second moment is large for r close to 2, and the higher moments are large for n close to r.

One characteristic of long-tail RV X is that its MGF

$$E[e^{sX}] \to \infty \quad \text{for all } s > 0.$$

A RV X is said to be light tailed (or short tailed) if $E[e^{sX}] < \infty$ for some $s > 0$.

Exponential distribution is short tailed.

One feature of heavy-tailed distribution is that its moments tend to be large. For properties of heavy-tailed distributions, see Sigman (1999).

Other standard long-tail distributions are lognormal and Weibull.

Consider the lognormal distribution defined as follows. If

$$U = (\log X - \zeta)/\sigma$$

is distributed as a standard normal variate, then the distribution of X is said to be lognormal (Johnson and Kotz, 1970). Its nth central moment

$$\mu_n' = E[X^n] = E[\exp n(\zeta + U\sigma)] = \exp\left(n\zeta + \frac{1}{2}n^2\sigma^2\right).$$

This increases very rapidly with n, and thus, $E[e^{sX}] \to \infty$. Lognormal distribution is heavy tailed.

Analyses of network traffic data indicate the existence of the long-tail nature of associated distributions, and long-tail service-time distributions are observed to occur in real-life situations. Such behavior may have considerable impact on performance measures of related queueing models. One such measure is waiting-time tail probabilities of a system in steady state. Some analyses of actual data indicate that long-tail interarrival times have much less impact on this measure than those of long-tail service-time distributions (as has been reported in some studies).

Alternative definitions of heavy-tail distributions are also given.

8.6.2.1 Subexponential distributions

Denote n fold convolution of $F(x)$ with itself by $F^{n*}(x)$. For $n = 2$

$$F^{2*}(x) = \int_0^x F(x - y)dF(y)$$

and so on. Denote the CDF by

$$F_c^{n*}(x) = 1 - F^{n*}(x). \tag{8.6.2}$$

Definition. The RV having DF $F(.)$ is called subexponential if $F_c(x) > 0, x \geq 0$ and for all $n \geq 2$,

$$\lim_{x \to \infty} \frac{F_c^{n*}(x)}{F_c(x)} = n. \tag{8.6.3}$$

It can be shown that if the above holds for some $n \geq 0$, then it holds for all $n \geq 2$.

It suffices therefore to consider a RV as subexponential if the above holds for $n = 2$.

For all practical purposes, the property of subexponentiality can be regarded as equivalent to $E[e^{sX}] \to \infty$ as for all $s > 0$. In fact, the class of subexponential distributions is a subclass of the class of heavy-tailed distributions.

The following theorem on the tail probabilities of the waiting of a general queueing system applies when the service time has such a distribution.

Pake's Theorem. *In a $G/G/1$ queue with service time V having mean unity and $E[e^{sv}] \to \infty$ for all $s > 0$, the tail probability of the waiting (queueing) time W is given by*

$$P(W > x) = \frac{\rho}{1 - \rho} {}_R B_c(x),$$

where ${}_R B_c(x)$ is the stationary excess (residual) service time of the complementary service time.

A corollary to this (Abate et al., 1994) is the following.

$$\text{If } E(V) = 1, \qquad P(V > x) = \alpha_r x^{-r} \quad \text{for } r > 1, \quad \text{then}$$
$$P(W > x) = \frac{\rho}{1 - \rho} \left(\frac{\alpha_r}{r - 1} \right) x^{-(r-1)} \quad \text{as } x \to \infty.$$

Notes:

(1) A RV V with $P(V > x) \sim \alpha_r x^{-r}$ is heavy tailed. (α_r might not depend on r; it could be a slowly varying function of x as $x \to \infty$.)

(2) In both of these cases, the results depend on the interarrival time distribution only through its mean.

The corollary could be used for computing waiting-time tail probabilities. Abate *et al.* show that the above two formulas can be very poor approximations for certain x values of interest.

8.6.3 $M/G/1$ with heavy-tailed service time

Studies on the $M/G/1$ model with heavy-tailed service-time distribution have been undertaken to find the tail probabilities of the waiting time. We recall that for the $M/G/1$ queue, the LST of the steady-state waiting (queueing) time distribution is given by the Pollaczek-Khinchin formula:

$$W^*(s) = \frac{1-\rho}{1-(\lambda/s)(1-B^*(s))}, \tag{8.6.4}$$

where $B^*(s)$ is the LST of the service-time distribution V. In terms of the residual service-time (stationary excess) distribution, the above can be expressed as

$$W^*(s) = \frac{1-\rho}{1-\rho\ {}_RB^*(s)}, \tag{8.6.5}$$

where ${}_RB^*(s) = (1-B^*(s))/sE(V)$ is the LST of the residual service time. (See Section 6.3.11a,b.)

Thus, in order to obtain $W^*(.)$ in an explicit form for an $M/G/1$ queue, it is necessary to have the LST of the service time V (or of the residual service time). Then by inversion of $W^*(.)$ analytically the waiting-time distribution can be found—or even when analytic inversion is cumbersome and does not yield results in explicit form, numerical inversion techniques could be employed (see, for example, Abate and Whitt (1992, 1995, 1996)).

The search is on for heavy-tail distributions having explicit LST to fit service-time distribution. Attention has been given to this topic, and a number of papers have recently been devoted to this.

Standard heavy-tail distributions, like Pareto, Weibull, and lognormal, do not have LST in explicit form. Heavy-tail distributions having LST in explicit form have been obtained through mixtures of distributions and employed as service-time distributions in queueing models. We discuss some below.

8.6.4 Pareto mixture of exponential (PME) distribution

Abate *et al.* (1994) search for a family of CDF having tail behavior as Pateto's—that is,

$$F_c(x;r) = \left(\frac{a}{x}\right)^r, \quad a>0, \quad x \geq a \tag{8.6.6}$$

having mean $ra/(r-1)$, which equals 1 for $a = (r-1)/r, (r > 1$ to have necessarily positive mean). An alternative form of the CDF is

$$F_c(x;r) = \left(\frac{r-1}{r}\right)^r x^{-r}, \quad x \geq (r-1)/r.$$

The PDF is given by

$$f(x; r) = r\left(\frac{r-1}{r}\right)^r x^{-(r+1)}, \quad x \geq (r-1)/r, \tag{8.6.7}$$

where r is an integer greater than 1, and by giving integral values to r, one gets a class of distributions. Its central moment (about the origin) is

$$\mu'_n = \frac{r}{r-n}\left(\frac{r-1}{r}\right)^n, \quad 1 \leq n < r. \tag{8.6.8}$$

Suppose that V is an exponential RV with mean y and that y has a Pareto distribution. Then one gets, through this mixture, a Pareto-Mixture of Exponential (PME) distribution having PDF

$$\begin{aligned} g(x; r) &= \int_{(r-1)/r}^{\infty} \{y^{-1}e^{-x/y}\} f(y; r)dy \\ &= \int_0^{r/(r-1)} r\left(\frac{r-1}{r}\right)^r v^r e^{-xv}\, dv \end{aligned} \tag{8.6.9}$$

(changing the variable from y to $v = 1/y$).

The nth central moment of this mixture distribution is given by

$$m'_n = \int_0^{\infty} E(V^n) f(y; r)dy = n!\mu'_n, \tag{8.6.10}$$

so that $m'_1 = 1$ and $m'_2 = 2\mu'_2$. For r close to 2, m'_2 is large. Thus, this distribution that has mean equal to 1 and large second moment is a heavy-tail distribution. The LST of the distribution is given by

$$\begin{aligned} g^*(s; r) &= \int_0^{\infty} e^{-sx} g(x; r)dx \\ &= r\left(\frac{r-1}{r}\right)^r \int_0^{r/(r-1)} v^r \left\{\int_0^{\infty} e^{-(s+v)x}\, dx\right\} dv \\ &= r\left(\frac{r-1}{r}\right)^r \int_0^{r/(r-1)} \frac{v^r}{s+v}\, dv. \end{aligned} \tag{8.6.11}$$

For integral r, by expanding, the integral can be put in the form

$$\begin{aligned} g^*(s; r) = &\sum_{i=1}^{r} (-1)^{r-i} \frac{r}{i}\left(\frac{r-1}{r}\right)^{r-i} s^{r-i} \\ &+ (-1)^r r\left(\frac{r-1}{r}\right)^r s^r \ln\left(1 + \frac{r}{(r-1)s}\right). \end{aligned} \tag{8.6.12}$$

For small $|s|$, the above formula (in powers of s) can be used for computation of $g^*(s; r)$.

For large $|s|$, the above can be put in the alternative form (in powers of $1/s$)

$$g^*(s;r) = \sum_{i=0}^{\infty}(-1)^i \frac{r}{r+1+i}\left(\frac{r}{r-1}\right)^{i+1} s^{-(i+1)}. \qquad (8.6.13)$$

An alternative form of $g^*(s;r)$ for $r = n+1/2$, when n is integral, is also given. One gets a long-tail distribution with explicit LST.

8.6.4.1 Waiting-time asymptotics

For an $M/G/1$ queue with service time having mean 1, let

$$W_c^*(s) \equiv \int_0^{\infty} e^{-sx} P(W > x)dx$$

be the LST of the CDF of the waiting time W. We have

$$\begin{aligned} W^*(s) &= \frac{1-\rho}{1-\rho_R B^*(s)} = \frac{1-\rho}{1-\rho\frac{(1-B^*(s))}{s}}, \quad \text{since } E(V)=1 \\ &= \frac{1}{\frac{1}{1-\rho} - \frac{\rho}{1-\rho}\left(\frac{1-B^*(s)}{s}\right)} \\ &= \frac{1}{1+\frac{\rho}{1-\rho}\left(1-\frac{1-B^*(s)}{s}\right)} = \frac{1}{1+\frac{\rho}{1-\rho}(1-{}_R B^*(s))} \\ &= \sum_{k=0}^{\infty}(-1)^k \left\{\frac{\rho}{1-\rho}(1-{}_R B^*(s)\right\}^k. \end{aligned}$$

Thus,

$$\begin{aligned} W_c^*(s) &= [1-W^*(s)]/s \\ &= \sum_{k=1}^{\infty}(-1)^{k+1}\left(\frac{\rho}{1-\rho}\right)^k \{1-{}_R B^*(s)\}^k s^{-1}. \end{aligned} \qquad (8.6.14)$$

Note: For PME service-time distribution with $r = 2$ (in an $M/G/1$ queue), Abate and Whitt obtain $P(W > x)$ by term by inversion of (8.6.14). Refer to Abate and Whitt (1994) for further results.

8.6.5 Gamma mixture of Pareto (GMP) distribution

Boxma and Cohen (1998b) discuss a mixture distribution having a heavy tail.

Consider an RV T_θ having a Pareto distribution of the second kind (also called the Lomax distribution) having CDF

$$B_c(t,\theta) = \delta\left(\frac{\theta}{\theta+t}\right)^{\nu}, \quad t \geq 0,\ \theta > 0,\ 0 < \delta \leq 1,\ 1 < \nu < 2 \qquad (8.6.15)$$

having mean $E(T_\theta) = \delta(\frac{\theta}{\nu-1})$; $E(T_\theta^2)$ is large for $\nu < 2$.

Consider that the parameter θ of the above distribution is itself distributed as a gamma RV having PDF

$$h(\theta) = \frac{a^{2-\nu}\theta^{1-\nu}e^{-a\theta}}{\Gamma(2-\nu)}, \quad a > 0, \quad 1 < \nu < 2 \tag{8.6.16}$$

having mean $\frac{2-\nu}{a}$.

The RV T obtained by mixing the Pareto distribution with a gamma distribution has DF

$$F(t) = P\{T \leq t\} = \int_0^\infty h(\theta)[1 - B_c(t,\theta)]d\theta. \tag{8.6.17}$$

T has mean

$$\begin{aligned} E(T) &= \int_0^\infty h(\theta)E(T_\theta)d\theta \\ &= \frac{\delta}{a}\frac{2-\nu}{\nu-1} = \beta \text{ (say)} \end{aligned} \tag{8.6.18}$$

and $E(T^2)$ is infinite. T has a heavy-tail distribution.

Denote the LST of T_θ and T by $B^*(s,\theta)$ and $B^*(s)$, respectively. The stationary excess distribution (forward recurrence time or residual lifetime) of T has LST

$${}_RB^*(s) = \frac{1 - B^*(s)}{s\beta} = \int_0^\infty h(\theta)\frac{1 - B^*(s,\theta)}{s\beta}d\theta.$$

Writing $\omega = a/s$, we get

$${}_RB^*(s) = \frac{\omega}{\omega - 1}\left[1 - \frac{1}{2-\nu}\frac{\omega^{2-\nu} - 1}{\omega - 1}\right]. \tag{8.6.19}$$

Thus, T could be fitted as a service-time distribution as it has heavy-tail and the residual service-time distribution has LST in explicit form. For a Poisson arrival single-server queue having T as service-time distribution, one can employ the Pollaczek-Khinchin formula (as given in 9.3.2).

As $E(T) \to \infty$, for $\nu = 1$ and $E(T) \to 0$, for $\nu = 2$, the cases $1 < \nu < 2$ are considered relevant. We consider here a particular case.

Particular case: $\nu = 1.5$. Then

$$\beta = E(T) = \frac{\delta}{a} \quad \text{and}$$

$$\begin{aligned} {}_RB^*(s) &= \frac{\omega}{\omega - 1}\left[1 - \frac{2}{\sqrt{\omega} + 1}\right] \\ &= \frac{1}{(1 + 1/\sqrt{\omega})^2} = \frac{a}{(\sqrt{s} + \sqrt{a})^2} \quad \text{(writing } a/s \text{ for } \omega), \end{aligned}$$

so that

$$B^*(s) = 1 - \frac{s\delta}{a}\,{}_R B^*(s)$$
$$= 1 - \delta\frac{a}{(\sqrt{s}+\sqrt{a})^2}. \tag{8.6.20}$$

Now we use the result. The Laplace transform inverse of

$$\frac{1}{\sqrt{s}+d} \text{ is } \frac{1}{\sqrt{\pi t}} - de^{d^2 t}\,\text{erfc}(d\sqrt{t}) \quad \text{where erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-u^2}\,du$$

(called complementary error function). Using the Laplace transform inversion formula, one gets

$$B(t) = 1 + \delta\left[\frac{2}{\sqrt{\pi}}\sqrt{at} - (1+2at)e^{at}\,\text{erfc}(\sqrt{at})\right], \quad t > 0$$
$$B(0+) = 1 - \delta. \tag{8.6.21}$$

Thus, the distribution of T and its LST can be found in explicit form.

Using the Pollaczek-Khinchin formula (from 6.3.11b), the LST of the waiting-time distribution is found to be

$$W^*(s) = (1-\rho)\left[1 + \frac{\rho a}{(\sqrt{s}+\sqrt{a})^2 - \rho a}\right]$$
$$= (1-\rho) + \frac{1}{2}(1-\rho)\sqrt{a\rho}$$
$$\left[\frac{1}{(\sqrt{s}+\sqrt{a}) - \sqrt{\rho a}} - \frac{1}{(\sqrt{s}+\sqrt{a}) + \sqrt{\rho a}}\right]. \tag{8.6.22}$$

Inversion of the Laplace transform of (8.6.22) gives the waiting time DF in explicit form, (for $t > 0$)

$$W(t) = 1 - \left(\frac{1+\sqrt{\rho}}{2}\right)\sqrt{\rho}e^{(1-\sqrt{\rho})^2 at}\,\text{erfc}((1-\sqrt{\rho})\sqrt{at})$$
$$+ \left(\frac{1-\sqrt{\rho}}{2}\right)\sqrt{\rho}e^{(1+\sqrt{\rho})^2 at}\,\text{erfc}((1+\sqrt{\rho})\sqrt{at}) \tag{8.6.23}$$

in terms of the complementary error function erfc(.).

Thus, we get the service-time distribution and the waiting-time distribution in explicit form for PME distribution employed as a heavy-tail service-time distribution (for the particular case $\nu = 1.5$).

Notes:

(1) Boxma and Cohen (1998b) also consider the asymptotics of waiting-time distribution.

(2) They also find explicit expression of the service-time distribution $B(t)$ for the general case $1 < \nu < 2$, and also the asymptotics of the waiting-time distribution.

(3) They also indicate that numerical computations can be carried by numerical inversion of the Laplace transform.

The heavy-tail distributions PME and GMP have been obtained by using the Pareto (itself a heavy-tail) distribution in a mixture, either as a mixing distribution or as a distribution of a parameter of another distribution.

There could be other types of mixture giving long-tail distribution with the facility of application of the Pollaczek-Khinchin formula. One such type considered by Abate and Whitt (1999b) is given next.

8.6.6 Beta mixture of exponential (BME) distribution

Such mixtures have been studied by Abate and Whitt (1999b). Consider that the RV T_y is exponential having PDF

$$f(x) = y^{-1}e^{-x/y}, \quad y > 0, \quad x \geq 0,$$

and let the mean y be distributed as a beta RV (of the first kind) having PDF

$$b(p, q; y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 \leq y \leq 1, \quad p, q > 0. \tag{8.6.24}$$

The RV beta mixture of exponential (BME) has PDF

$$v(p, q; x) \equiv \int_0^1 y^{-1}e^{-x/y} b(p, q; y)dy, \quad x \geq 0. \tag{8.6.25}$$

Considering beta distribution of the second kind with PDF

$$b_2(p, q; y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1+y)^{-(p+q)}, \quad y \geq 0, \quad p, q > 0$$

one can get another mixture (BME-2) distribution.

The nth central moment of the BME distribution is given by

$$\begin{aligned} m_n' \equiv m_n'(p, q) &= \int_0^1 E\left[T_y^n\right] v(p, q; x)dx \\ &= n! \frac{\Gamma(p+n)\Gamma(p+q)}{\Gamma(p)\Gamma(p+q+n)} \\ &= n! \frac{(p)_n}{(p+q)_n}, \end{aligned}$$

where $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)} = a(a+1)\ldots(a+n-1)$, $n \geq 1$ with $(a)_0 = 1$. Thus

$$m_1'(p,q) = \frac{p}{p+q}, \qquad m_2'(p,q) = \frac{2p(p+1)}{(p+q)(p+q+1)}$$

$$m_3'(p,q) = \frac{6p(p+1)(p+2)}{(p+q)(p+q+1)(p+q+2)} \quad \text{and}$$

$$\frac{m_3'(p,q)}{m_1'(p,q)m_2'(p,q)} = 3\left(\frac{p+q}{p}\right)\left(\frac{p+2}{p+q+2}\right). \tag{8.6.26}$$

Denote the LSH by $F(p,q)$; then $\frac{\partial F}{\partial q} > 0$ and $\frac{\partial F}{\partial p} < 0$, so that F is increasing in q and decreasing in p. That is, for fixed mean and variance, $m_3'(p,q)$ is increasing in q and decreasing in p. In general, for fixed mean and variance, $m_n'(p,q)$ is increasing in q and decreasing in p for all $n \geq 3$. The moments m_n' of the BME distribution are large for large q and n. This implies that BME distribution is long tailed.

Now let us consider the LST $v^*(p,q;s)$ of BME distribution

$$\begin{aligned} v^*(p,q;s) &= \int_0^\infty e^{-st} v(p,q;t)\,dt \\ &= \int_0^\infty \sum_{n=0}^{\infty} \frac{(-st)^n}{n!} v(p,q;t)\,dt \\ &= \sum_{n=0}^{\infty} (-s)^n \frac{(p)_n}{(p+q)_n}. \end{aligned} \tag{8.6.27}$$

The series is the well-known Gauss hypergeometric series $F(1,p;p+q;-s)$ with radius of convergence 1. For $|s| < 1$, the series on the RHS could be computed numerically.

An alternative form of expansion considered by them is to use the result

$$F(1,p;p+q;-s) = \frac{1}{1+s} F(1,q;p+q;s/(s+1)),$$

$$\text{whence} \quad v^*(p,q;s) = \frac{1}{1+s} v^*(q,p;-s/(1+s))$$

$$= \sum_{n=0}^{\infty} \frac{s^n}{(1+s)^{n+1}} \frac{(q)_n}{(p+q)_n}. \tag{8.6.28}$$

Inversion of the LST $v^*(p,q;s)$ is facilitated by the fact that the Laplace transform of the Laguerre function

$$l_n(t) = e^{-t} \sum_{k=0}^{n} \binom{n}{k} \frac{(-t)^k}{k!}$$

is

$$\frac{s^n}{(1+s)^{n+1}}.$$

Thus, for $p, q > 0$

$$v(p,q;t) = \sum_{n=0}^{\infty} l_n(t) \frac{(q)_n}{(p+q)_n}. \tag{8.6.29}$$

Notes:

(1) Abate *et al.* also investigate BME-2 distribution and several properties of the BME, BME-2 distributions.

(2) They also consider a gamma mixture of exponential (GME) distribution. When the parameter y of an exponential distribution is distributed as gamma distribution, then one gets a gamma mixture of exponential (GME) distribution with PDF $u(p,q;t)$ such that

$$\lim_{q\to 0} \frac{u(p,q;t/q)}{q}$$

is expressible in terms of Bessel function. They also consider the LST of GME.

Also note that if the mean $\mu(=\frac{1}{y})$ of the exponential distribution is distributed as a gamma distribution, the mixture results in Pareto distribution of the second kind (Lomax distribution).

8.6.7 A class of heavy-tail distributions

In a recent paper, Abate and Whitt (1999a) consider a class of distributions that can be used in modeling service-time distribution having long-tail and LST in explicit form. Let V be a RV having PDF $g(t)$ and LST $g^*(s)$ given by

$$g^*(s) = \int_0^{\infty} e^{-st} g(t)dt = 1 - \frac{s}{(1+\sqrt{s})(\mu+\sqrt{s})}. \tag{8.6.30}$$

It follows that $E(V) = 1/\mu$ and that all higher moments $E(V^n)$, $n \geq 2$ are infinite. Then the distribution of V is heavy tailed and can be taken as the service-time distribution of a queueing system. The LST of the residual service time of distribution V is given by

$${}_Rg^*(s) = \frac{1-g^*(s)}{sE(V)} = \frac{\mu}{(1+\sqrt{s})(\mu+\sqrt{s})}.$$

The case $\mu = 1$ corresponds to the particular case $\nu = 1.5, \delta = 1$, considered in Section 8.6.4.1. Here the case $\mu \neq 1$ is considered; for $\mu \neq 1$,

$${}_Rg^*(s) = \frac{\mu}{1-\mu}\left(\frac{1}{\mu+\sqrt{s}} - \frac{1}{1+\sqrt{s}}\right). \tag{8.6.31}$$

Inverting the LST one gets

$$ {}_R g(t) = \frac{\mu}{1-\mu}\{e^t \operatorname{erfc}(\sqrt{t}) - \mu e^{\mu^2 t} \operatorname{erfc}(\mu\sqrt{t})\}, \tag{8.6.32} $$

which is the PDF of the stationary excess distribution of V (residual service-time distribution).

Using the Pollaczek-Khinchin formula, the LST of the waiting time can be obtained. Denote by ${}_c W^*(s)$ the LST of the *conditional* waiting time that there is a positive wait; one then gets (see Note 2, Section 6.3.2.1).

$$ \begin{aligned} {}_c W^*(s) &= \frac{(1-\rho)\,{}_R g^*(s)}{1-\rho\,{}_R g^*(s)} \\ &= (1-\rho)\mu\left\{\frac{1}{s+(1+\mu)\sqrt{s}+(1-\rho)\mu}\right\} \\ &= \frac{(1-\rho)\mu}{\nu_1-\nu_2}\left\{\frac{1}{\nu_2+\sqrt{s}} - \frac{1}{\nu_1+\sqrt{s}}\right\}, \end{aligned} \tag{8.6.33} $$

where ν_1, ν_2 are the roots of $s + (1+\mu)\sqrt{s} + (1-\rho)\mu = 0$.

Again the LST of the complementary distribution function CDF

$$ \begin{aligned} W_c^*(s) &= \frac{\rho}{s}[1 - {}_c W^*(s)] \\ &= \frac{\rho}{\nu_1-\nu_2}\left[\frac{\nu_1}{\sqrt{s}(\nu_2+\sqrt{s})} - \frac{\nu_2}{\sqrt{s}(\nu_1+\sqrt{s})}\right]. \end{aligned} \tag{8.6.34} $$

By inverting the LST, one gets

$$ W_c(t) = P(W > t) = \frac{\rho}{\nu_1-\nu_2}\left[\nu_1 \Psi\left(\nu_2^2 t\right) - \nu_2 \Psi\left(\nu_1^2 t\right)\right], $$

where $\Psi(t) = e^t \operatorname{erfc}(\sqrt{t})$, which tends to $1/\sqrt{\pi t}$ as $t \to \infty$.

Thus, the asymptotics of the waiting time are found.

Notes:

(1) The distribution with the LST given in (8.6.30) is, in fact, a mixture distribution. For this and other discussions, see Abate and Whitt (1999a).

(2) So far, heavy-tail distributions have been shown as being mixtures of distributions. It is also shown that hyperexponential distribution having CDF

$$ H_c(t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t}, \quad p_i \geq 0, \quad \sum p_i = 1 $$

(and LST $H_c^*(s) = \sum_{i=1}^{k}(p_i\lambda_i)/(\lambda_i + s)$) could also be used as an approximation of a long-tail distribution. (See Feldmann and Whitt, 1998). They also develop an algorithm for construction of suitable approximating distributions for a class of heavy-tail distributions.) For further details, see the papers mentioned herein and the relevant references.

8.6.8 Long-range dependence

Temporal high variability is another characteristic of traffic processes encountered now; this characteristic was absent with voice traffic alone.

Statistical analyses reveal that correlations of traffic processes are generally found to exhibit a hyperbolic decay, indicating long-range dependence rather than an exponential decay with short-range dependence. Mandelbrot pointed out the analogous problem of estimating the length of a coastline that needs to have a wide range of length scales, and the accurate measurement of the length depends on the sensitivity of the yardstick.

If $r(k)$ is the autocorrelation function of a time series associated with a traffic process, then a short-range dependence implies exponentially decaying $r(k)$ such that

$$r(k) \sim a(k)\rho^k \quad \text{as } k \to \infty, \ 0 < \rho < 1,$$

where $a(k)$ is a slowly varifying function at infinity. Short-range dependence allows for limited burstiness or fluctuations.

Traffic processes are said to be long-range dependent if the $r(k)$ of the associated time series is such that $r(k) \sim b(k)k^{2H-2}$, as $k \to \infty, \frac{1}{2} < H < 1$, where $b(k)$ is slowly varying at infinity, and H is (called) the *Hurst parameter* (which is commonly used as a measure of long-range dependence in a time series). Exponential decay occurs in case of short-range dependence, and hyperbolic decay occurs in case of long-range dependence, indicating that the associated traffic process is statistically *self-similar.*

The phenomenon of self-similarity has been a subject of recent studies (see, for example, Hampel *et al.* (1986), Willinger *et al.* (1995, 1996), and the references therein). Self-similar stochastic processes, introduced by Kolmogorov, were brought to the attention of probabilists and statisticians by Mandelbrot and his coworkers. Self-similar processes are discussed in books, for example, by Hampel *et al.* (1986) and Peitgen *et al.* (1992).

These are fractal features, involving many length and time scales, which were nonexistent so long in queueing context and now characterize traffic processes. These new trends in traffic processes point to a new direction of research in queueing theory, called fractal queueing theory. This appears to be an emerging area and has been engaging attention.

Erramilli *et al.* (1997) indicate that there is considerable scope for further research in this area of fractal queueing, and they also draw attention to some open issues.

We wish to conclude by mentioning this development, which is a very new and rapidly evolving area of research, with many of the papers still in the preprint stage. Interested readers are referred to the references cited for further details.

The new ideas indicated would be of use in practical traffic-related studies, especially in telecommunication systems.

Problems and Complements

8.1. Consider an $M/G/1$ system. Let $W(t)$ be the virtual delay, and let a busy period start at t_o with the arrival of a customer whose service time v has DF $B()$ with LST $B^*()$ and moments b_k. The delay process $\{W(t),\ t \geq 0\}$ can be approximated as a diffusion process with infinitesimal mean and variance given by

$$c = \rho - 1$$
$$D^2 = \lambda b_2.$$

Let $f(t,x)$ be the first passage of time from 0 to x of the Wiener process with infinitesimal mean $-c$ and variance D^2, and let $f(t)$ be the PDF of the busy period for the approximate process. Show that

$$f(t,x) = x(2\pi D^2 t^3)^{-1/2} \exp\left\{\frac{-(x+ct)^2}{2D^2 t}\right\} \quad \text{and}$$

$$f(t) = (2\pi D^2 t^3)^{-1/2} \int_0^\infty x \exp\left\{-\frac{(x+ct)^2}{2D^2 t}\right\} dB(x).$$

For an $M/M/1$ queue

$$f(t) = \{\mu\rho(\pi^2\rho t)^{-1/2}\} \exp\left\{-\frac{(1-\rho)\mu t}{4\rho}\right\}$$
$$-\left\{\frac{(3\rho-1)}{2}\right\}\mu \exp\{(2\rho-1)\mu t\}\Phi_c\left[\left\{\frac{(3\rho-2)}{2}\right\}\left(\frac{\mu t}{\rho}\right)^{1/2}\right],$$

where Φ_c is the complementary DF of a standard normal variate,

$$\Phi_c(x) = (2\pi)^{-1/2} \int_x^\infty e^{-y^2/2}\, dy.$$

For an $M/D/1$ queue with constant service time d

$$f(t) = (2\pi\lambda t^3)^{-1/2} \exp\left[\frac{-\{d+(\rho-1)t\}^2}{2\rho\, dt}\right].$$

(Heyman, 1974)

8.2. $M/G(Q,\infty)/1$: average long-run cost. Assume that the input is Poisson with rate λ, and the service times are IID random variables, denoted by

v, with moments $b_k = E(v^k)$. The service rule is general bulk service with infinite capacity having a minimum of Q in a batch. Let K be the startup cost and h be the waiting cost per customer per unit time. Let R be the interval between two renewals—that is, between the epochs of two successive service initiations, and let Y be the cost between two renewals. Show that

$$E(R) = b_1 + (1/\lambda) \sum_{n=0}^{Q-1} (Q - n) P_n,$$

where P_n is the probability that exactly n customers arrive during a service period, and

$$E(Y) = K + h\lambda b_2 + (h/2\lambda) \sum_{n=0}^{Q-1} (Q^2 - Q - n^2 + n) P_n.$$

Hence, find the mean waiting time in an $M/G(Q,\infty)/1$ queue (Weiss, 1979).

8.3. $M/G/1$ under N-policy with zero start-up time.
Suppose that demand for service arises in accordance with a Poisson process with rate λ. The service times are IID random variable, with DF B having first *two* finite moments $b_1 = \mu$ and b_2. The facility starts (instantaneously) only when N units are present after a busy period with exhaustive service terminates (i.e., the server becomes idle).

A fixed setup cost $K > 0$ is incurred every time the facility is reopened, and a holding cost $h(>0)$ per unit time per unit present is also incurred for every unit present in the system. This is an $M/G/1$ system under N-policy with zero start-up time. N is the control parameter.

(a) Show that the long-run fraction of time the service facility is busy equals ρ independently of the control parameter N.

(b) Show that, with probability 1, the long-run average cost per unit time is given by

$$\frac{\lambda(1-\rho)K}{N} + h\left\{\rho + \frac{\lambda^2 b_2}{2(1-\rho)} + \frac{N-1}{2}\right\}$$

and that the average cost is minimal for one of the two integers nearest to

$$N = \sqrt{\left\{\frac{2\lambda(1-\rho)K}{h}\right\}}.$$

(Tijms, 1986)

8.4. $M/G/1$ queue under T-policy with zero start-up time.
Here the facility is controlled in a different manner. Every time the server becomes idle after exhaustive service, the service facility is utilized for other work for fixed length of time T, and then the facility is reactivated only when there is at least one unit (i.e., multiple-server vacation with fixed vacation time T). Suppose that K is the fixed setup cost, h is the holding cost per unit time per unit present, and start-up time is zero.

(a) Defining a cycle C as the interval between two consecutive epochs at which a vaction period starts, show that

$$E(C) = \frac{T}{1-\rho}.$$

(b) Show that with probability 1, the long-run average cost per unit time is given by

$$\frac{K(1-\rho)}{T} + h\left\{\rho + \frac{\lambda^2 b_2}{2(1-\rho)} + \frac{\lambda T}{2}\right\}$$

and that the average cost per unit time is mimimal for

$$T = \sqrt{\frac{2(1-\rho)K}{h\lambda}}.$$

(Tijms, 1986)

8.5. $M/G/1$ system under N-policy and general start-up time.
Suppose that instead of zero start-up time, the start-up times are IID random variables with common DF $D(.)$ with mean u and LST $D^*(.)$ The system can be in three different types of states: I (turned off), S (turned-on with server doing preservice work) (as soon as queue size builds up to N after a turned-off period I), and B (server rendering service to customers that starts as soon as turned-on period is over and terminates when all present are served, leaving none in the queue) (exhaustive service). Let $p_{n,i}$ denote the steady-state probability that there are n in the system, given that the state is $i, i \in \{I, S, B\}$. Let $P_I(z)$, $P_s(z)$, and $P_B(z)$ be the corresponding PGFs. Show that

$$p_{n,I} = p_{0,I} = k(1-\rho), \quad 0 \le n \le N-1, \quad \text{and}$$
$$P_s(z) = p_{0,I}\frac{z^N\{D^*(\lambda - \lambda z) - 1\}}{(z-1)}.$$

Using the decomposition property, show further that (with $k = 1/(N + \lambda u)$ and $B^*()$ as the LST of service-time distribution)

$$P_B(z) = \frac{k\rho(1-\rho)\{z^N D^*(\lambda - \lambda z) - 1\}B^*(\lambda - \lambda z)}{z - B^*(\lambda - \lambda z)}$$

For busy-period distribution, see Section 6.4.5 (Medhi and Templeton, 1992).

8.6. Kella (1989) considers an $M/G/1 - V_m$ system with a threshold policy. Here the server goes on taking vacations of random length V until the first instance after a vacation where he finds at least N units in the system (N is a pressigned positive integer) and begins service immediately and exhaustively (until none is left in the system). It is assumed that the durations of the vacations $\{V\}$ are independent of the arrival process and the service times.

Consider the busy cycle C_N composed of the total vacationing time (server idle period) and the total server busy period B_N. Denote by M_N the number of vacations that the server takes, and by b_i the probability that there are i arrivals during a vacation $i = 0, 1, 2, \ldots$

(1) Show that the following recursive relations hold.

$$\text{(a)}\quad E\{C_N\} = \frac{E(V)}{1-\rho} + \sum_{k=0}^{N} b_{N-k} E\{C_k\}, \quad E\{C_0\} = 0$$

$$\text{(b)}\quad E\{V_N\} = \left[E(V) + \sum_{k=0}^{N-1} b_{N-k} E\{V_k\} \right] \Big/ [1 - V^*(\lambda)], \; E\{V_0\} = 0$$

$$\text{(c)}\quad E\{M_N\} = \left[1 + \sum_{k=0}^{N-1} b_{N-k} E\{M_k\} \right] \Big/ [1 - V^*(\lambda)], \; E\{M_0\} = 0$$

Verify that

$$\frac{E\{B_N\}}{E\{C_N\}} = \rho$$

$$\text{and} \quad E\{B_N\} = \frac{\rho}{1-\rho} E\{V_N\}.$$

Obtain for a standard $M/G/1 - V_m$ queue (as a particular case), the expressions for

$$E\{C_1\}, E\{V_1\} \text{ and } E\{M_1\}.$$

(2) Show further that the steady-state probability π_i that the number of customers at the commencement of vacations is given by

$$\pi_i = \frac{E\{C_{i+1}\} - E\{C_i\}}{E\{C_N\}}, \quad i = 0, 1, \ldots, N-1$$

$$\text{and} \quad \sum_{i=0}^{N-1} \pi_i = 1.$$

Note that the PGF of the distribution of system size L_N can be obtained writing

$$\zeta(z) = \sum_{i=0}^{N-1} \pi_i z^i$$

for $\zeta(z)$ in (8.3.21). It can be found that

$$E\{L_N\} = E\{L\}_{M/G/1-V_m} + \left[N - \frac{\sum_{i=1}^{N} E\{C_i\}}{C_N} \right]$$

(the last term is due to the N-policy—that is, N-policy related and equals 0 when $N = 1$ (standard $M/G/1 - V_m$).
For details see Kella (1989), who also considers the control policy and obtains the optimum threshold policy for the average and total cost criteria.

8.7. Reliability analysis with breakdown and repair. Li *et al.* (1997) consider an $M/G/1$ queue with server breakdowns where service time X is general. They consider the situation that the server has an exponential lifetime with mean $1/\alpha$ and is subject to breakdown when serving units. The server is then sent for repairs, with repair time Y having mean $1/\beta$. As soon as the server is fixed (repaired), it begins serving units, starting with the unit whose service was interrupted due to breakdown. After each service completion, the server takes a Bernoulli vacation V with probability q or continues to serve other units in queue with probability $p = 1 - q$ [$p = 1$ corresponds to exhaustive service]. The server always takes a vacation V, as soon as the system is empty. At the end of it, the server starts service, if there is any unit waiting, and otherwise it waits for units to arrive (single vacation when the system is empty). Suppose that the arrival rate is λ and the service rate is μ.

Show that the steady-state probabilities that the system is busy, empty (free), under repair, or on vacation ρ, P_E, P_R, P_V, respectively, are given by

$$\rho = \lambda/\mu = a$$

$$P_E = \frac{(1-c)V^*(\lambda)}{V^*(\lambda) + \lambda p E(V)}$$

$$P_R = \frac{\lambda\alpha}{\mu\beta}$$

$$\text{and} \quad P_V = q\lambda E(V) + \frac{\lambda p(1-c)E(V)}{V^*(\lambda) + \lambda p E(V)},$$

$$\text{where} \quad c = \left(\frac{\lambda}{\mu}\right)\left(1 + \frac{\alpha}{\beta}\right) + \lambda q E(V).$$

It may also be shown that

$$E\{W\} = \frac{1}{\mu}\left(1 - \frac{\alpha}{\beta}\right) + \frac{\lambda}{2(1-c)}\left[\frac{\alpha}{\mu}E(Y^2) + \left(1 + \frac{\alpha}{\beta}\right)^2 E(X^2) + qE(V^2) + \frac{2q}{\mu}\left(1 + \frac{\alpha}{\mu}\right)E(V)\right] + \frac{\lambda p E(V^2)}{2[V^*(\lambda) + \lambda p E(V)]}$$

Obtain this result by heuristic argument. (See Section 8.3.9.)
They further consider availability $A(t)$ of the system at time t (either, the server is busy or the system is empty) and show that, in steady state

$$A = \lim_{t\to\infty} A(t) = \frac{\lambda}{\mu} + \frac{(1-c)V^*(\lambda)}{V^*(\lambda) + \lambda p E(V)}.$$

See Li *et al.* (1997) for details as well as for transient analysis.
As a particular case, with $\alpha = 0, \beta \to \infty, p = 1$, one gets the corresponding results for an $M/G/1 - V_s$ system where

$$c = \frac{\lambda}{\mu} = \rho$$

$$P_E = (1-\rho)\frac{V^*(\lambda)}{[V^*(\lambda) + E(V)]}$$

$$P_R = 0$$

$$P_V = (1-\rho)\cdot\frac{\lambda E(V)}{V^*(\lambda) + E(V)}$$

$$E(W) = \frac{1}{\mu} + \frac{\lambda b^{(2)}}{2(1-\rho)}$$

and $A = \rho + P_E$.

Interpret the above results.

8.8. Consider an $M/G/1/N$ queue with multiple vacation and exhaustive service. Assume that the duration of vacation is of constant length V. Then the probabilitities π_j that there are j customers in the system at an arbitrary point of time are given by

$$\pi_j^*(V) = \frac{\pi_j(V)(1 - e^{-\lambda V})/\lambda}{V\pi_0(V) + E(B)(1 - e^{-\lambda V})} \qquad j = 0, 1, \ldots, N-1$$

$$= 1 - \frac{(1 - e^{-\lambda V})/\lambda}{V\pi_0(V) + E(B)(1 - e^{-\lambda V})}, \qquad j = N,$$

where $\pi_j(V)$ and $\pi_j^*(V)$ denote, respectively, π_j and π_j^* those with constant vacation duration V (as are given by the solution of (8.3.46) and (8.3.47) with $N = K$).

Thus (taking limit as $V \to 0$), we get,

$$\pi_j^*(0) = \frac{\pi_j(0)}{\pi_0(0) + \rho}, \quad j = 0, 1, \ldots, N-1$$
$$= 1 - \frac{1}{\pi_0(0) + \rho}, \quad j = N,$$

which give the probability distribution of the system size in an $M/G/1/N$ queue (without vacation) (Frey and Takahashi, 1997b).

References and Further Reading

Abate, J., Choudhury, G. L., and Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems* **16**, 311–338.

Abate, J., and Whitt, W. (1992). The Fourier series method for inverting transforms of probability distributions. *Queueing Systems* **10**, 5–88.

Abate, J., and Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA J Computing* **7**, 36–43.

Abate, J., and Whitt, W. (1996). An operational calculus for probability distributions via Laplace transforms. *Adv. Appl. Prob.* **28**,75–113.

Abate, J., and Whitt, W. (1999a). Explicit $M/G/1$ waiting-time distributions for a class of long-tail service-time distributions. *Opns. Res. Lett.* **25**, 25–31.

Abate, J., and Whitt, W. (1999b). Modeling service-time distributions with non-exponential tails: beta mixtures of exponentials. *Commn. Stat.-Stochastic Models.* **15 (3)**, 517–546 (with 35 references).

Adler, R., Feldman, R., and Taqqu, M. S. (Eds.) (1998). *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy-Tailed Distributions*, Birkhäusser, Boston.

Aissani, A., and Artalejo, J. R. (1998). On the single server retrial queue subject to breakdowns. *Queueing Sys.* **30**, 309–321.

Ali, O. M. E., and Neuts, M. F. (1984). A service system with two stages of waiting and feedback of customers. *J. Appl. Prob.* **21**, 404–423.

Altiok, T. (1987). Queues with group arrivals and exhaustive service discipline. *Queueing Systems* **2**, 307–320.

Artalejo, J. R. (1998). Retrial queues with a finite number of sources. *J. Korean Math. Soc.* **35**, 503–525.

Artalejo, J. R. (1999a). Accessible bibliography on retrial queues. *Math. and Comp. Modeling* **30**, 1–6.

Artalejo, J. R. (1999b). Classified bibliography of research on retrial queues: Progress in 1990–99. *TOP*, 7 # 2, 187–211 (contains a list of 148 references).

Baba, Y. (1986). On the $M^X/G/1$ queues with vacation time. *Opns. Res. Lett.* **5**, 93–98.

Bell, C. E. (1971). Characterization and computation of optimal policies for operating on $M/G/1$ queueing system with removable server. *Opns. Res.* **23**, 571–574.

Bell, C. E. (1975). Turning off a server with customer present. Is this any way to run on $M/G/c$ queue with removable servers? *Opns. Res.* **23**, 571–574.

Benes, V. E. (1956). On queueing with Poisson arrivals. *Ann. Math. Stat.* **28**, 670–677.

Berman, O., Larson, R. C., and Shiu, S. S. (1985). Optimal server location in a network operating as an $M/G/1$ queue. *Opns. Res.* **33**, 746–771.

Bertsekas, D., and Gallager, R. (1992). *Data Networks*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Borovkov, A. A. (1984). *Asymptotic Methods in Queueing Theory*, Wiley, New York.

Boxma, O. J., and Cohen, J. W. (1998a). The Single Server Queue: Heavy Tails and Heavy Traffic in *Self-Similar Network Traffic and Performance Evaluation* (Eds. K. Park and W. Willinger), Wiley, New York.

Boxma, O. J., and Cohen, J. W. (1998b). The $M/G/1$ queue with heavy tailed service time distributions. *IEEE J. on Selected Areas in Communications*, **16 #5**, 749–763.

Boxma, O. J., and Yechiali, U. (1997). An $M/G/1$ queue with multiple types of feedback and gated vacations, *J. Appl. Prob.* **34**, 773–784.

Chae, K. C., and Lee, H. W. (1995). $M^X/G/1$ vacation models with N-policy: heuristic interpretation of the mean waiting time. *J. Opl. Res. Soc.* **46**, 258–264.

Chaiken, J. M., and Dermont, N. P. (1978). A patrol car allocation model: background, capabilities, and algorithms. *Mgmt. Sci.* **24**, 1280–1300.

Chatterjee, U., and Mukherjee, S. (1990). $GI/M/1$ queue with server vacation. *J.Opl. Res. Soc.* **41**, 83–87.

Choudhury, G. (2000). An $M^X/G/1$ queueing system with a setup period and a vacation period. *Queueing Systems* **36**, 23–38.

Choudhury, G. (2001). A batch arrival queue with a vacation time under single vacation policy. *Comp. & Opns. Res.* (To appear).

Chung, K. L., and Williams, R. (1990). *An Introduction to Stochastic Integration*, Birkhäuser, Boston.

Cooper, R. B. (1970). Queues served in cyclic order: waiting times. *Bell Sys. Tech. J.* **49**, 399–413.

Cox, D. R. (1972). Regression models and life table. *J.R.S.S.* **B34**, 187–200.

Crabill, T. B., Gross, D., and Magazine, M. J. (1977). A classified bibliography of research on optimal design and control of queues. *Opns. Res.* **25**, 219–232.

Deb, R. K., and Serfozo, R. F. (1973). Optimal control for batch service queue. *Adv. Appl. Prob.* **5**, 340–361.

Doshi, B. T. (1978). Optimal control of the service rate in an $M/G/1$ queueing system. *Adv. Appl. Prob.* **10**, 662–701.

Doshi, B. T. (1985). A note on stochastic decomposition in a $GI/G/1$ queue with vacation or set-up times. *J. Appl. Prob.* **22**, 419–428.

Doshi, B. T. (1986). Queueing systems with vacations—a survey. *Queueing Systems* **1**, 29–66.

Doshi, B. T. (1990). Single Server Queues with Vacations in *Stochastic Analysis of Computer & Comm. Systems* (Ed. H. Takagi), 217–265, Elsevier Science, Publ. B.V., Amsterdam.

Dshalalow, J. H. (Ed.) (1995). *Advances in Queueing Theory: Methods and Open Problems*, CRC Press, Boca Raton, FL.

Dyer, M. E., and Proll, L. G. (1977). On the validity of marginal analysis for allocating servers in $M/M/c$ queues. *Mgmt. Sci.*, **23**, 1019–1022.

Erramilli, A., Narayan, O., and Willinger, W. (1997). Fractal Queueing Models in *Frontiers in Queueing: Models and Applications in Science and Engineering* (Ed. J. H. Dshalalow), 245–269, CRC Press, Boca Raton, FL.

Erramilli, A., and Willinger, W. (1993). Fractal Properties of Packet Traffic Measurements in *Proc. of St. Petersburg Reg. ITC Seminar St. Petersburg*, 144–158.

Falin, G. I. (1990). A survey of retrial queues. *Queueing Syst.* **7**, 127–168 (contains a list of 109 references).

Falin, G. I., and Templeton, J. G. C. (1997). *Retrial Queues*, Chapman & Hall, London.

Federgruen, A., and Green, L. (1986). Queueing systems with service interruptions. *Opns. Res.* **34**, 752–768.

Feldmann, A., and Whitt, W. (1998). Fitting mixtures of exponentials to long tail distributions to analyze network performance models. *Perf. Ev.* **31**, 245–279 (with 59 references).

Fredericks, A. A. (1982). A class of approximations for the waiting time distribution in $GI/G/1$ queueing system. *Bell. System. Tech. J.* **61**, 295.

Frey, A., and Takahashi, Y. (1997a). A note on an $M/G/1/N$ queue with vacation time and exhaustive service discipline. *Opns. Res. Lett.* **33**, 1117–1129.

Frey, A., and Takahashi, Y. (1997b). Explicit Solutions for the $M/G/1/N$ Finite Capacity Queues With and Without Vacation Time in *Teletraffic Contributions for the Information Age* (Ed. V. Ramaswamy and P. E. Wrath), 507–516, Elsevier, Amsterdam.

Frey, A., and Takahashi, Y. (2000). An $M^x/G/1/N$ queue with close down and vacation times. *J. Appl. Math. & Stoc. An.* (To appear).

Fricker, C. (1986). Etude d'une file $GI/G/1$ à service autonome (avec vacances du serveur). *Adv. Appl. Prob.* **18**, 283–286.

Fuhrmann, S. W. (1984). A note on the $M/G/1$ queue with server vacations. *Opns. Res.* **32**, 1368–1373.

Fuhrmann, S. W., and Cooper, R. B. (1985). Stochastic decompositions in an $M/G/1$ queue with generalized vacations. *Opns. Res.* **33**, 1117–1129.

Gaver, D. P., Jr. (1962). A waiting line with interrupted service, including priorities. *J.R.S.S.* **B24**, 73–90.

Gaver, D. P., Jr. (1968). Diffusion approximations and modes for certain congestion problems. *J. Appl. Prob.* **5**, 607–623.

Gelenbe, E. (1979). Probabilistic models of computer systems. Part II : Diffusion approximations, waiting times and batch arrivals. *Acta Informatica* **12**, 285–303.

Gelenbe, E., and Iasnogorodski, R. (1980). A queue with server of walking type (autonomous service). *Ann. Inst. Henri Poincare* **XVI**, 63–73.

Gelenbe, E., and Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems*, Academic Press, London.

Grassman, W. (1983). The convexity of the mean queue size of the $M/M/c$ queue with respect to the traffic intensity. *J. Appl. Prob.* **20**, 916–919.

Green, L. (1984). A multiple dispatch queueing model of police patrol operations. *Mgmt. Sci.* **30**, 653–664.

Halachmi, B., and Franta, W. R. (1978). A diffusion approximation to the multiserver queue. *Mgmt. Sci.* **24**, 522–529; Erratum, 1448.

Hall, P. (1985). Heavy traffic approximations for busy period in an $M/G/\infty$ queue. *Stoch. Proc. & App.* **19**, 259–269.

Hampel, F. R., Ronchetti, E. M., Rousseauv, P. J., and Stahel, W. A. (1986). *Robust Statistics*, Wiley, New York.

Hanschke, T. (1997). Explicit formulas for the characteristics of the $M/M/2/2$ queue with repeated attempts. *J. Appl. Prob.* **24**, 486–494.

Harel, A., and Zipkin, P. (1987). Strong convexity results for queueing systems. *Opns. Res.* **35**, 405–418.

Harris, C. M., and Marchal, W. G. (1988). State dependence in $M/G/1$ server-vacation models. *Opns. Res.* **36**, 560–565.

Harrison, M. J. (1985). *Brownian Motion and Stochastic Flow Systems*, Wiley, New York.

Heckman, J., and Singer, B. (1984). A method of minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**, 271–320.

Heyman, D. P. (1968). Optimal control policies for $M/G/1$ queueing systems. *Opns. Res.* **16**, 362–382.

Heyman, D. P. (1974). An approximation for the busy period of the $M/G/1$ queue using a diffusion model. *J. Appl. Prob.* **11**, 159–169.

Heyman, D. P. (1975). A diffusion model approximation for the $GI/G/1$ queue in heavy traffic. *Bell Syst. Tech. J.* **54**, 1637–1646.

Heyman, D. P. (1977). The T-policy for the $M/G/1$ queue. *Mgmt. Sci.* **23**, 775–778.

Heyman, D. P., and Sobel, M. J. (1982). *Stochastic Models in Operations Research*, McGraw-Hill, New York.

Hillier, F. S., and Lieberman, G. J. (1967). *Introduction to Operations Research*, Holden-Day, San Francisco.

Huang, C. C., Brumelle, S. L., and Sawaki, K. (1977). Optimal control for multiserver queueing systems under periodic reviews. *Nav. Res. Log. Q.* **24**, 127–135.

Iglehart, D. L. (1965). Limit diffusion approximations for the many server queues and the repairmen problem. *J. Appl. Prob.* **2**, 429–441.

Iglehart, D. L., and Whitt, W. (1970). Multiple channel queues in heavy traffic. I & II. *J. Appl. Prob.* **2**, 150–177, 355–369.

Ignall, E., and Kolesar, P. (1974). Optimal dispatching of an infinite capacity shuttle: control at a single terminal. *Opns. Res.* **22**, 1003–1024.

Jacob, M. J., and Madhusoodanan, T. P. (1987). Transient solution for a finite capacity $M/G(a, b)/1$ queueing system with vacations to the server. *Queueing Systems* **2**, 381–386.

Jagerman, D. L., Melamed, B., and Willinger, W. (1997). Stochastic Modelling of Traffic Processes in *Frontiers in Queueing* (Ed. J. H. Dshalalow), 271–320, CRC Press, Boca Raton, FL.

Johnson, N. L., and Kotz, S. (1970). *Continuous Univariate Distributions,* 1 & 2, Wiley, New York.

Jonin, G. L., and Sedol, Y. Y. (1970). Investigation of telephone systems in presence of repeated calls. *Latvian Math. Yearbook* **7**, 71–83, Riga (in Russian).

Karaesmen, F., and Gupta, S. M. (1996). The finite capacity $GI/M/1$ queue with server vacations. *J. Opl. Res. Soc.* **47**, 817–828.

Keilson, J., Cozzolino, J., and Young, H. (1968). A service system with unfilled request repeated. *Opns. Res.* **16**, 1126–1137.

Keilson, J., and Servi, L. D. (1987). The dynamics of an $M/G/1$ vacation model. *Opns. Res.* **35**, 575–582.

Keilson, J., and Servi, L. D. (1989). Blocking probability for $M/G/1$ vacation systems with occupancy level dependent schedules. *Opns. Res.* **37**, 134–140.

Kella, O. (1989). The threshold policy in the $M/G/1$ queue with server vacations. *Nav. Res. Log.* **36**, 111–123.

Kimura, T. (1983). Diffusion approximation for an $M/G/m$ queue. *Opns. Res.* **31**, 304–321.

Kimura, T., and Ohsone, T. (1984). A diffusion approximation for an $M/G/m$ queue with group arrivals. *Mgmt. Sci.* **30**, 381–388.

Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Proc. Camb. Phill. Soc.* **57**, 902–904.

Kingman, J. F. C. (1962a). On queues in heavy traffic. *J.R.S.S.* **B24**, 383–392.

Kingman, J. F. C. (1962b). Some inequalities for the queue $GI/G/1$. *Biometrika* **49**, 315–324.

Kingman, J. F. C. (1964). The Heavy Traffic Approximation in the Theory of Queues in *Proceedings of Symposia on Congestion Theory* (Eds. W. L. Smith and W. E. Wilkinson), 137–169, University of North Carolina Press, Chapel Hill, NC.

Kingman, J. F. C. (1970). Inequalities in the theory of queues. *J.R.S.S.* **B32**, 102–110.

Kingman, J. F. C. (1980). Queue disciplines in heavy traffic. *Math. Opns. Res.* **7**, 262–271.

Kitaev, M. Y., and Rykov, V. V. (1995). *Controlled Queueing Systems,* CRC Press, New York.

Kleinrock. L. (1976). *Queueing Systems* Vol. II, *Computer Applications,* Wiley, New York.

Kobayashi, H. (1974). Application of the diffusion approximation to queueing networks: Part I: Equilibrium queue distributions. Part II: Non equilibrium distributions and applications to computer modeling. *J. Assoc. Comput. Mach.* **21**, 316–328, 459–469.

Kollerström, J. (1974). Heavy traffic theory for queues with several servers I. *J. Appl. Prob.* **11**, 544–552.

Kosten, L. (1967). The Custodian Problem in *Queueing Theory: Recent Developments and Applications* (Ed. R. Cruon), pp. 65–70. English Press Ltd.

Krishnamoorthy, A., and Ushakumari, P. V. (2000). A queueing system with single arrival bulk service and single departure. *Math. and Computer Modeling* **31**, 99–108.

Kulkarni, V. G. (1983). On queueing systems with retrials. *J. Appl. Prob.* **20**, 380–389.

Kulkarni, V. G., and Liang, H. M. (1997). Retrial Queues Revisited in *Frontiers in Queueing* (Ed. J. H. Dshalalow), 18–34, CRC Press, Boca Raton, FL.

Kushner, H. J. (2001). *Heavy Traffic Analysis of Controlled Queueing and Communications Networks,* Springer, New York.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica* **47**, 939–956.

Lazar, A. A. (1983). Optimal flow control of a class of queueing networks in equilibrium. *IEEE Trans. Ant. Cont.* **AC-28**, 1001–1007.

Lee, H. L., and Cohen, M. A. (1983). A note on the convexity of performance measures of $M/M/c$ queueing systems. *J. Appl. Prob.* **20**, 920–923.

Lee, H. W., Lee, S. S., Park, Jeong, O. K., Chae, K. C. (1994). Analysis of the $M^x/G/1$ queue with N-policy and multiple vacations. *J. Appl. Prob.* **31**, 476–496.

Lee, H. W., Lee, S. S., Yoon S. H., and Chae, K. C. (1995). Batch arrival queue with N-policy and single vacation. *Comp. Ops. Res.* **22**, 173–189.

Lee, H. W., Yoon, S. H., and Lee, S. S. (1996). A continuous approximation for batch arrival queues with threshold. *Comp. Ops. Res.* **23**, 299–308.

Lee, T. T. (1984). $M/G/1/N$ queue with vacation time and exhaustive service discipline. *Opns. Res.* **32**, 774–785.

Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1993). On the Self-Similar Nature of Ethernet Traffic in *Proc. of ACM Sigcomm '93*, San Francisco, 183–193. Extended version of the paper. (1994). *IEEE/ACM Trans on Networking*, 1–15.

Lemoine, A. (1975). Limit theorems for generalized single server queues: the exceptional system. *SIAM J. Appl. Math.* **28**, 596–606.

Levy, H., and Kleinrock, L. (1986). A queue with starter and a queue with vacations: delay analysis by decompositions. *Opns. Res.* **34**, 426–436.

Levy, Y., and Yechiali, U. (1975). Utilization of idle time in an $M/G/1$ queueing system. *Mgmt. Sci.* **22**, 202–211.

Levy, Y., and Yechiali, U. (1976). $M/M/s$ queues with server vacations. *INFOR* **14**, 153–163.

Li, Wei, Shi, Dinghua, and Chao Xiuli. (1997). Reliability analysis of $M/G/1$ queueing systems with server breakdowns and vacations. *J. Appl. Prob.* **34**, 546–555.

Lippman, S. A. (1975). Applying a new device in the optimization of exponential queuing systems. *Opns. Res.* **23**, 687–710.

Loris-Teghem, J. (1988). Vacation policies for an $M/G/1$ type queueing system with finite capacity. *Queueing Systems* **3**, 41–52.

Lucantoni, D. M., Meier-Hellstern, and Neuts, M. F. (1990). A single-server queue with server vacations and a class of non-renewal arrival processes. *Adv. Appl. Prob.* **22**, 676–705.

Mandelbrot, B. B. (1965). Self-similar error clusters in communication systems and the concept of conditional stationarity. *IEEE Trans. Comm. Tech.* COM-13, 71–90.

Marchal, W. G. (1976). An approximate formula for waiting time in single-server queues. *AIIE Trans.* **8**, 473.

Marchal, W. G. (1987). An empirical extension of the $M/G/1$ heavy traffic approximation. *Annals O.R.* **8**, 93–101.

Medhi, J. (1984a). Bulk Service Queueing Models and Associated Control Problems in *Statistics: Applications and New Directions*, 369–377, Indian Statistical Institute, Calcutta.

Medhi, J. (1984b). *Recent Developments in Bulk Queueing Models*, Wiley Eastern, New Delhi.

Medhi, J. (1997). Single-Server Queueing System with Poisson Input in *Advances in Combinatorial Methods and Applications to Probability and Statistics* (Ed. N. Balakrishanan), 317–388, Birkhäusser, Boston.

Medhi, J. (2001). Teletraffic Theory: Some Recent Developments in *Advances in Probability Theory and Stochastic Processes* (Eds. A. Krishnamoorthy, N. Raju, and V. Ramaswamy), 113–124, Notable Publications, NJ.

Medhi, J., and Templeton, J. G. C. (1992). A Poisson input queue under N- policy and with general start-up time. *Comp. & Opns. Res.* **19**, 35–41.

Miller, B. I. (1968). Finite state continuous time Markov decision processes with finite planning horizon. *SIAM. J. Control.* **6**, 266–280.

Mitchell, B. (1973). Optimal service rate selection in an $M/G/1$ queue. *SIAM. J. Appl. Math.* **24**, 19–30.

Mitrani, I. L., and Avi-Itzhak, B. (1968). A many server queue with service interruptions. *Opns. Res.* **16**, 628–638.

Neuts, M. F., and Lucantoni, D. M. (1979). A Markovian queue with N servers subject to breakdowns and repairs. *Mgmt. Sci.* **25**, 849–861.

Neuts, M. F., and Ramalhoto, M. F. (1984). A service model in which the server is required to search for customers. *J. Appl. Prob.* **21**, 157–166.

Neuts, M. F., and Rao, B. M. (1990). Numerical investigation of a multiserver retrial model. *Queueing Syst.* **7**, 169–189.

Newell, G. F. (1971). *Applications of Queueing Theory* (2nd ed. 1982), Chapman & Hall, London.

Nobel, R. (1998). *Hysteretic and Heuristic Control of Queueing Systems* (Doctoral thesis, Vrije Universiteit, Amsterdam).

Nobel, R. D., and Tijms, H. C. (1999). Optimal control for a $M^x/G/1$ queue with two service modes. *Euro. J. Opnl. Res.* **113**, 610–619.

Ohno, K., and Ichiki, K. (1997). Computing optimal policies for controlled tandem queueing systems. *Opns. Res.* **35**, 121–126.

Pakes, A. (1972). A $GI/M/1$ queue with a modified service mechanism. *Ann. Inst. Stat. Math.* **24**, 589.

Pakes, A. (1975). On the tails of waiting time distributions. *J. Appl. Prob.* **25**, 132–141.

Parikh, S. C. (1977). On a fleet sizing and allocation problem. *Mgmt. Sci.* **23**, 972–977.

Parlar, M. (1984). Optimal dynamic service rate control in time dependent $M/M/S/N$ queues. *Int. J. Systems. Sci.* **15**, 107–118.

Paxson, V. (1994). Growth trends in wide-area TCP connections. *IEEE Network* **8** (4), 8–17.

Peitgen, O., Jeurgens, H., and Saupe, D. (1992). *Chaos and Fractals: New Frontiers of Science*, Springer-Verlag, New York.

Pourbabai, B., and Sonderman, D. (1986). Server utilization factors in queueing loss systems with ordered entry and heterogeneous servers. *J. Appl. Prob.* **23**, 236–242.

Powell, W. B. (1981). *Stochastic Delays in Transportation Terminals: New Results in the Theory and Application of Bulk Queues*, Ph.D. Dissertation, MIT, Cambridge, MA.

Powell, W. B., and Humblet, P. (1986). The bulk service queue with a general control strategy: theoretical analysis and new computational procedure. *Opns. Res.* **34**, 267–275.

Prabhu, N. U. (1974). Stochastic control of queueing systems. *Nav. Res. Log. Q.* **21**, 411–418.

Ramaswamy, R., and Servi, L. D. (1988). The busy period of the $M/G/1$ vacation model with a Bernoulli schedule. *Stoch. Models* **4**, 507–521.

Reiser, M. (1982). Performance evaluation of data communication systems. *Proc. of IEEE* **70**, 171–196 (includes a list of 142 references).

Reiser, M., and Kobayashi, H. (1974). Accuracy of the diffusion approximation for some queueing systems. *IBM J. Res. Devel.* **18**, 110–124.

Rolfe, A. J. (1971). A note on the marginal allocation in multi-server facilities. *Mgmt. Sci.* **17**, 656–658.

Ross, S. M. (1970a). Average cost semi-Markov decision processes. *J. Appl. Prob.* **7**, 649–656.

Rue, R. C., and Rosenshine, M. (1981). Some properties of optimal control policies for entries to an $M/M/1$ queue. *Nav. Res. Log. Q.* **28**, 225–232.

Sauer, C. H., and Chandy, K. (1981). *Computer System and Performance Modelling*, Prentice-Hall, Engelwood Cliffs, NJ.

Serfozo, R. (1981). Optimal control of random walks, birth and death processes and queues. *Adv. Appl. Prob.* **13**, 61–83.

Shanthikumar, J. G. (1988). On stochastic decomposition in $M/G/1$ queues with generalized server vacations. *Opns. Res.* **36**, 566–569.

Shanthikumar, J. G., and Yao, D. D. (1986). The effects of increasing service rates in closed queueing network. *J. Appl. Prob.* **23**, 474–483.

Shanthikumar, J. G., and Yao, D. D. (1987). Optimal server allocation in a system of multi-server stations. *Mgmt. Sci.* **33**, 1173–1191.

Sigman, K. (1999). A primer on heavy-tailed distributions. *Queueing Systems* **33**, 261–275.

Sivazlian, B. D. (1979). Approximate optimal solution for a D-policy in an $M/G/1$ queueing system. *AIIE Trans.* **11**, 341–343.

Sobel, M. J. (1969). Optimal average cost policy for a queue with start-up and shut down costs. *Opns. Res.* **17**, 145–162.

Sobel, M. J. (1974). Optimal Operation of Queues in *Mathematical Methods of Queueing Theory* (Ed. A. B. Clarke), pp. 231–261, Springer-Verlag, Berlin, New York.

Stepanov, S. N. (1983). *Numerical Methods for Calculation for Systems with Repeated Calls*, Nauka, Moscow (in Russian).

Stidham, S. (1985). Optimal control of admissions to a queueing system. *IEEE Trans Automat. Control AC-30*, 705–713.

Stidham, S., Jr., and Prabhu, N. U. (1974). Optimal Control in Queueing Systems in *Mathematical Methods of Queueing Theory* (Ed. A. B. Clarke), 263–294, Springer-Verlag, Berlin, New York.

Sunaga, T., Kondo, E., and Biswas, S. K. (1978). An approximation method using continuous models for queueing problems. *J. Opns. Res. Soc. Japan* **21**, 29–44.

Szarkowicz, D. S., and Knowles, T. W. (1985). Optimal control of an $M/M/s$ queueing system. *Opns. Res.* **33**, 644–660; Err. 34, 184.

Takagi, H. (1991, 1993). *Queueing Analysis: A Foundation of Performance Evaluation*, Vol. 1 *Vacation and Priority System*, Vol. 2 *Finite Systems*. North Holland, Amsterdam.

Takagi, H. (1992a). Analysis of an $M/G/1/N$ queue with multiple server vacation. *J. Opnl. Res. Soc. Japan*, **35**, 300–315.

Takagi, H. (1992b). Time dependent process of $M/G/1$ vacation models with exhaustive service. *J. Appl. Prob.* **29**, 418–429.

Takine, T., and Hasegawa, T. (1992). A generalization of the decomposition property in the $M/G/1$ queue with server vacations. *Opns. Res. Lett.* **12**, 97–99.

Teghem, J., Jr. (1986). Control of the service process in a queueing system. *Euro. J. Opnl. Res.* **30**, 141–158.

Teghem, J., Jr. (1987). Optimal control of a removable server in an $M/G/1$ queue with finite capacity. *Euro. J. Opnl. Res.* **31**, 358–367.

Tian, N., Zhang, D., and Cao, C. (1989). $GI/M/1$ queue with exponential vacations. *Queueing Syst.* **5**, 331–344.

Tijms, H. C. (1986). *Stochastic Modelling and Analysis: A Computational Approach*, Wiley, Chichester, UK.

Tijms, H. C. (1994). *Stochastic Modelling and Analysis: An Algorithmic Approach*, Wiley, New York.

Tu, H. Y., and Kumin, H. (1983). A convexity result for a class of $GI/G/1$ queueing systems. *Opns. Res.* **31**, 948–950.

Van-Nunen. J. A. E. E., and Puterman, M. L. (1983). Computing optimal control limits of $GI/M/s$ queueing systems with controlled arrivals. *Mgmt. Sci.* **29**, 725–734.

Weber, R. R. (1980). On the marginal benefit of adding servers to $G/GI/m$ queues. *Mgmt. Sci.* **26**, 946–951.

Weber, R. R. (1983). A note on waiting times in single-server queues. *Opns. Res.* **31**, 950–951.

Weiss, A. (1995). An introduction to large deviations for communication networks. *IEEE J. on Sel. Areas in Comm.* **13**, 938–952.

Weiss, H. J. (1979). The computation of optimal control limits for a queue with batch services. *Mgmt. Sci.* **25**, 320–328.

Weiss, H. J. (1981). Further results on an infinite capacity shuttle with control at a single terminal. *Opns. Res.* **29**, 1212–1217.

Whitt, W. (1974). Heavy Traffic Limit Theorems for Queues: A Survey in *Mathematical Methods of Queueing Theory* (Ed. A. B. Clarke), 307–350, Springer-Verlag, Berlin, New York.

Whitt, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT & T Bell Lab. Tech. J.* **63**, 689–708.

Whitt, W. (1985). The best order for queues in series. *Mgmt. Sci.* **31**, 475–487.

Whitt, W. (1999). Predicting queueing delays. *Mgmt. Sci.* **45**, 870–888.

Willekens, J. L., and Teugels, J. L. (1992). Asymptotic expansions for waiting time probabilities in $M/G/1$ queue with long-tailed service time. *Queueing Systems* **10**, 295–312.

Willinger, W., and Paxson, V. (1998). Where mathematics meets the Internet. *Notices of the American Math. Soc.* **45#8**, 961–970.

Willinger, W., Taqqu, M. S., and Erramilli, A. (1996). A Bibliographical Guide to Self-Similar Traffic and Performance Modeling for Modern High Speed Networks in *Stochastic Networks: Theory and Applications* (Eds. F. P. Kelly, S. Zachary, and I. Ziedins) (contains a list of 420 references), 339–366, Clarendon Press, Oxford.

Willinger, W., Taqqu, M. S., Leland, W. E., and Wilson, D. V. (1995). Self-similarity in high speed packet traffic: analysis and modelling of Ethernet traffic measurements. *Statistical Science* **10**, 67–85.

Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, NJ.

Yadin, M., and Naor, P. (1963). Queueing systems with a removable service station. *Opnl. Res. Qrly.* **14**, 393–405.

Yang, T., Posner, M. J. M., and Templeton, J. G. C. (1990). The $M/G/1$ retrial queue with nonpersistent customers. *Queueing Syst.* **7**, 209–218.

Yang, T., and Templeton, J. G. C. (1987). A survey on retrial queues. *Queueing Syst.* **2**, 203–233. (Contains a list of 65 references.)

Yao, D. D. (1985). Refining the diffusion approximation for the $M/G/m$ queue. *Opns. Res.* **33**, 1266–1277.

Yao, D. D. (1986). Convexity properties of the over-flow in an ordered entry system with heterogeneous servers. *Opns. Res. Lett.* **5**, 145–147.

Abbreviations and Symbols

ASTA	Arrivals see time averages
BCMP	Baskett, Chandy, Muntz, Palacios
CCDF; CDF	Complementary distribution function
C-K, C.K.	Chapman-Kolmogorov
DF	Distribution function
DFR	Decreasing failure rate
e	Column vector with each of its elements equal to unity
$E(X)$, $E\{X\}$	Expectation of the random variable X
FCFS	First-come, first-served
FCLT	Functional Central Limit Theorem
FIFO	First-in, first-out
FLLN	Functional-law-of-large-numbers
F^*G	Convolution of two independent distributions with DFs F and G
F^{n*}, $F^{(n)*}$	Convolution of n IID random variables with common distribution F
$F^{*(k)}(s)$	kth-derivative of $F^*(s)$ (with $F^{*(1)}(s) \equiv F^{*'}(s)$)
$f^*(s)$, $\tilde{f}(s)$	Laplace transform of $f(t)$
IID	Identically and independently distributed
iff	If and only if
IMLR	Increasing mean residual life
LAA	Lack of anticipation assumption

LCFS	Last-come/first-served
LHS	Left-hand side
LST	Laplace-Stielties Transform
LT	Laplace Transform
MGF	Moment generating function
MPBA	Multiple Poisson bulk arrivial
$\mathbf{0}$	Column vector with each of its elements equal to zero
$\mathbf{P} = (p_{ij})$	A matrix with elements p_{ij}
PASTA	Poisson arrivals see time averages
PDF	Probability density function
PGF	Probability generating function
P-K, PK	Pollaczek-Khinchin
$Pr(A)$, $P(A)$, $Pr\{A\}$, $P\{A\}$	Probability of the event A
RHS	Right-hand side
RV	Random variable
SCV	Squared coefficient of variation
s.d., SD	Standard deviation
SMP	Semi-Markov process
SUT	Start-up time or setup time
TPM	Transition probability matrix
var(X)	Variance of the random variable X
WRT	With respect to
$\boldsymbol{\alpha} = (a, \ldots, a_n)$	A vector with elements, $a_1, \ldots, a_n$

Properties of Laplace Transforms

$$L\{f(t)\} = \bar{f}(s) = \int_0^\infty e^{-st} f(t)dt \quad (t \geq 0)$$

one-to-one correspondence exists between $f(t)$ and $\bar{f}(s)$.

1. Linearity property

$$L\{a_1 f_1(t) + \cdots + a_k f_k(t)\} = a_1 \bar{f}_1(s) + \cdots + a_k \bar{f}_k(s)$$

2. Translation property

(i) $L\{e^{-at} f(t)\} = \bar{f}(s + a)$

(ii) $L\begin{Bmatrix} f(t-a), & t > a \\ 0, & t < a \end{Bmatrix} = e^{-as} \bar{f}(s)$

3. Change-of-scale property

$$L\{f(at)\} = \left(\frac{1}{a}\right) \bar{f}\left(\frac{s}{a}\right)$$

4. LT of derivatives

$$L\{f'(t)\} = s\,\bar{f}(s) - f(0)$$

and in general

$$L\{f^{(n)}(t)\} = s^n \bar{f}(s) - \sum_{i=1}^{n} s^{n-i} f^{(i-1)}(0)$$

for derivative $f^{(n)}(t)$ of order $n, n = 1, 2, \ldots$

5. LT of integrals

   (i) $L\left\{\int_0^t f(x)dx\right\} = \dfrac{\bar{f}(s)}{s}$

   (ii) $L\left\{\int_0^t \int_0^u f(x)dx\,du\right\} = \dfrac{\bar{f}(s)}{s^2}$

6. Convolution property

$$L\{f * g\} = L\left\{\int_0^t g(t-y)\,f(y)dy\right\} = \bar{f}(s)\bar{g}(s)$$

7. Limit property: Initial value property

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} s\,\bar{f}(s)$$

8. Limit property: Final value property

$$\lim_{t\to 0} f(t) = \lim_{s\to\infty} s\,\bar{f}(s)$$

9. Multiplication by power of t

$$L\{t^n f(t)\} = (-1)^n \bar{f}^{(n)}(s)$$

   (where $\bar{f}^{(n)}(.)$ denotes nth derivative of $\bar{f}(.)$)

10. Inversion Formula

$$L^{-1}\{\bar{f}(s)\} = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{ts}\,\bar{f}(s)ds = f(t)$$

Table of Laplace Transforms

$$\bar{f}(s) = \int_0^\infty e^{-st} f(t)\,dt$$

	$\bar{f}(s)$	$f(t)$
1.	$\dfrac{1}{s}$	1
2.	$\dfrac{1}{s^n}$	$\dfrac{t^{n-1}}{(n-1)!}, \quad n = 2, 3, \ldots$
3.	$\dfrac{1}{s^a}$	$\dfrac{t^{a-1}}{\Gamma(a)}, \quad a > 0$
4.	$\dfrac{1}{\sqrt{s}}$	$\dfrac{1}{\sqrt{\pi t}}$
5.	$\dfrac{1}{s-a}$	e^{at}
6.	$\dfrac{1}{(s-a)(s-b)}$	$\dfrac{1}{a-b}(e^{at} - e^{bt}), \quad a \neq b$
7.	$\dfrac{1}{(s-a)^n}$	$\dfrac{t^{n-1}e^{at}}{(n-1)!}, \quad n = 2, 3, \ldots$
8.	$\dfrac{a}{s+a}$	$ae^{-at}, \quad a > 0$

9.	$\dfrac{a}{s^2 + a^2}$	$\sin at$
10.	$\dfrac{s}{s^2 + a^2}$	$\cos at$
11.	$\left(\dfrac{a}{s + a}\right)^k$	$\dfrac{a^k t^{k-1} e^{-at}}{\Gamma(k)}, \quad k > 0,\ a > 0$
12.	$\dfrac{2as}{(s^2 + a^2)^2}$	$t \sin at$
13.	$\dfrac{s^2 - a^2}{(s^2 + a^2)^2}$	$t \cos at$
14.	$\ln\left(1 + \dfrac{1}{s}\right)$	$\dfrac{1 - e^{-t}}{t}$
15.	$\dfrac{s}{(s + a)(s + b)}$	$\dfrac{ae^{-at} - be^{-bt}}{a - b}, \quad a \neq b$
16.	$\dfrac{b}{(s + a)^2 + b^2}$	$e^{-at} \sin bt$
17.	$\dfrac{s}{(s^2 + a^2)(s^2 + b^2)}$	$\dfrac{\cos at - \cos bt}{b^2 - a^2}, \quad a^2 \neq b^2$
18.	e^{-as}	$f(t) = \delta(t - a) = 1, \quad t = a$ $= 0, \quad t \neq a$ (Dirac δ function located at a)
19.	$\dfrac{e^{-as}}{s}$	$f(t) = 0, \quad 0 < t < a$ $= 1, \quad a < t$ (unit step function)
20.	$\dfrac{e^{-as}}{s^2}$	$f(t) = 0, \quad 0 < t < a$ $= t - a, \quad a < t$
21.	$\dfrac{e^{1/s}}{s^{n+1}}$	$t^{n/2} I_n(2\sqrt{t})$
22.	$\dfrac{\{s - (s^2 - a^2)\}^n}{\sqrt{s^2 - a^2}}$	$a^n I_n(at), \quad n > -1$
23.	$\{s - (s^2 - a^2)\}^n$	$\dfrac{n}{t} I_n(at), \quad n = 1, 2, \ldots$

24. $\dfrac{\sqrt{s}}{s-a^2}$ $\quad \dfrac{1}{\sqrt{\pi t}} + ae^{a^2 t}\operatorname{erf}(a\sqrt{t})$

$$\left(\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,dt : \text{error function}\right)$$

25. $\dfrac{1}{\sqrt{s}+a}$ $\quad \dfrac{1}{\sqrt{\pi t}} - ae^{a^2 t}\operatorname{erfc}(a\sqrt{t})$

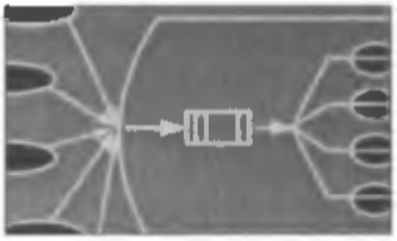
$$\left(\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\,dt,\right.$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{1}{\sqrt{\pi}}\,\Gamma\left(\frac{1}{2}, x^2\right),$$

$$\left.\text{where} \quad \Gamma(a,x) = \int_x^\infty e^{-t} t^{a-1}\,dt\right)$$

This Page Intentionally Left Blank

Index



A

Approximation
  diffusion, 389–397
  heavy traffic, 375–382
  mean waiting time, 353–355
Arrival pattern, 48
ASTA, 62
Asymptotic
  behavior, 383–384
  queue-length distribution, 386–388
Average
  customer, 54
  time, 54

B

Backlog, 391
Balance equations, 227–233
  flow, 227
  global, 232
  local, 233
Balking, 154
  and reneging, 154
Batch
  arrival, 30, 48
  service, 30, 49
BCMP, 238–240
Bernoulli
  feedback, 243–244, 302–303
  scheduled vacation, 408
Birth-death processes, 23–25, 81–83
  pure birth process, 23
  pure death process, 23
Blocking formula, *see* Erlang
Brownian motion
  definition, 383
  process (diffusion process), 383–386
Bulk queue, 174–217
  bulk-arrival system, 174–185
  bulk-service system, 185–209
  general bulk-service rule, 186
Burke's theorem, 53
Busy cycle, 285
Busy period, *see* specific models

C

Capacity of a system, 49
Central Limit Theorem,
  functional, 382
  for heavy traffic, 375–378
  of queueing theory, 378
Channel
  parallel, 49
Chapman-Kolmogorov (C-K)
  equation, 4, 15, 17, 22, 27, 137, 165–166, 171, 187
  forward and backward, 17
Characterization problem, 273
Compound Poisson process, 29
Computer communication system, 408
Concavity property, 93, 424

Congestion
  call, 107–108
  time, 107–108
Conservation
  equations, 227
  methods, 54
Control policy, 146
Convexity property, 324, 424
  of performance measures, 324
Conveyor system, 138
Cost-benefit analysis, 54
Customer
  average, 57
  conservation, 57
  primary and secondary, 398
  super, 296
Cycle time
  in cyclic exponential
  network, 245
Cyclic queue (model), 236–238

D

Decomposition
  property, 399
  result, 400, 406–407
  stochastic, 399
Delay
  busy period, 284–285
  busy period under N-policy, 285–289
  system, 49
Design and control of queues, 423–427
Diffusion
  approximation, 389–397
  equation, 385–386
  model, 383
  process, 383–386
Discipline, *see* Queue discipline
Distribution
  BME, 450–452
  Borel-Tanner, 284
  busy period, *see* Busy period of
    specific model
  compound Poisson, 29
  Coxian, 239, 245, 306
  Engset, 102, 107
  Erlang, 165–173
  GMP, 447–450
  lognormal, 443
  long-tail (heavy tail), 442–455
  Pareto, 442, 445, 447
  PME, 445–447
  short-tail (lean-tail), 442
  sub-exponential, 443–444
  Weibull, 445
D-Policy, 425

E

$E_k/M/1$, 170–174
Elementary return boundary, 387
Embedded Markov chain, 40
  technique for $M/G/1$ queue, 256–259
  technique for $M/G(a,b)/1$ queue, 304–306
Engset model
  delay model, 101–102
  loss model, 106–108
Entropy maximization, 31
Ergodicity, 5, 16
Ergodic system, 5
Ergodic theorem, 6–7
Erlang
  B formula (blocking/overflow/loss), 99–100
  C formula, 86
  delay formula, 86–89
  recursive algorithm, 99–100
  relation between B and C formulas, 100
  unit (erlang), 50

F

Feedback queues, 302–304
Finite input source, 101, 289–291
First entrance formula, 9
First passage time, 119
  distribution, 124–125
Fluid
  flow, 383
  model, 383
Folk theorem, 312
Forward equation, *see* Chapman-Kolmogorov
    equation
Fredericks' approximation, *see* $G/G/1$
Functional
  Central Limit Theorem, 383
  law of large numbers, 383

G

Gaussian process, 388
General relationships in queueing theory, 54–59
$G/G/1$, 339–371
  approximation of mean waiting time,
    353–355

asymptotic queue-length distribution, 387–388
bound of expected idle period, 360
bounds of mean waiting time, 360–368
diffusion approximation, 389–391
expected waiting time, 348–351
Fredericks' approximation, 379–382
generalization of P-K transform formula, 346–347
heavy traffic approximation, 379–382
interdeparture time (interval), 358–359
Kingman's approximation, 375–378
Lindley's integral equation, 341–342
output process, 358–360
virtual delay, 391–393
waiting-time distribution, 348–353
$G/G/c$
approximation of mean waiting time, 353–355
average number of busy channels, 57
diffusion approximation, 395–396
heavy traffic approximation, 382
output process, 360
$G/M/1$, 306–314
arrival epoch state probabilities, 306–309
expected duration of busy and idle periods, 313–314
general time system size, 309–311
system size at most recent arrival, 311
with vacation, 415
waiting-time distribution, 311–313
$G/M/c$, 319–322
heavy traffic approximation, 381–382
waiting-time distribution, 321–322
$G/M/c//m.$, 103
$G/M/m/m$, 97
$G^x/G/1$ model, 356–358
Gordon and Newell model, 233–236
Gordon and Newell theorem, 233–234
Group arrivals, *see* Batch arrival
Group service, *see* Batch service

H

Heavy-tail distribution, *see* Distribution
Heavy-traffic
approximation, 375–382
Central Limit Theorem/result, 383
distribution of waiting time, 375–382
Fredericks' approximation, *see* $G/G/1$
Kingman's approximation *see* $G/G/1$

I

Idle period, 38–39
Infinitesimal generator, *see* Markov chain
Input (or arrival) pattern, 48
Instantaneous return process, 394
Interdeparture interval, 72–74
interoverflow time, 144
Invariant measure, *see* Stationary distribution
Invariant (or insensitivity) property, 317–318
Inverse problem, 54

J

Jackson network, 225, 226–233
Jackson's theorem, 227–233

K

Khinchin, *see* Pollaczek-Khinchin
Kingman
approximation, 375–378
conjecture, 382
Kobayashi approximation, 390
Kolmogorov equation, 385

L

Lack of anticipation assumption (LAA), 60
Ladder process, 368
Laplace transform, 471–475
final value theorem of, 472
initial value theorem of, 472
numerical inversion of, 445
properties, 471–472
table of, 473–475
of waiting-time distributions, 343–344
Law of large numbers strong, 382
Level crossing argument, 405
Lifetime
past (or spent), 36–37
residual, 36–37
Lindley's integral equation, 341–343
Little's formula, 55–58
distributional form of, 58
Eilon's proof, 55–56
generalization of, 57–58
Load
carried, 50
offered, 50

Local balance equation, *see* Balance equations
Long-range dependence, 454–455, 505
Long-tail distributions, *see* Distribution
Loss networks, 241–242
Loss system, 49

M

Machine interference (repairmen) problem, 101, 289
Markov chain, 2–22
  aperiodic, 5
  classification of states and chains, 4–13
  continuous time, 14–25
  denumerable, 8–9
  derived, 32–34
  embedded, 40, 250, 256
  ergodicity property, 5
  ergodic theorems, 6–13
  essential, inessential, 4
  homogeneous, 4
  infinitesimal generator, 15–16
  invariant measure of a (stationary measure of a), 5–7
  irreducible, 5
  periodic, 5
  primitive, 5
  regular, 5
  sojourn time in, 14–15
  transience and recurrence, 9–10
  transition probability matrix, 3
  underlying on a Poisson process, 33
Markov process, 14, 32, 33
Markov renewal process , 39–41
Markovian Arrival Process, 324–326
Matrix
  canonical form, 5
  routing (switching, transfer), 232
  stochastic, 3
  transition density, 16
  transition probability matrix, 3
$M/D/1$, 261
Mean value analysis heuristic treatment, 416–423
Memoryless property, 28
Method of supplementary variable, 256, 267–274
$M/E_k/1$, 165–170
  equivalence with $M^k/M/1$, 178
$M/G/1$, 255–306
  bulk arrivals, *see* $M^x/G/1$,
  busy period, 277–284
  delay busy period, 284–288
  departure epoch system size, 260
  diffusion approximation, 390, 393
  embedded Markov chain technique, 256–259
  feedback, 302
  heavy traffic approximation, 379–381
  interdeparture interval, 334
  martingale approach, 274–275
  Pollaczek-Khinchine formula, 259–260, 267–274
  processor sharing, 245–246
  semi-Markov process approach, 274
  supplementary variable technique, 256, 267–274
  Takács integral equition, 277–279, 287
  waiting-time distribution, 261–267
  with a second optional channel, 275–276
  with vacations, 400–412
$M/G/1/K$, 292–294
  with vacation, 412–414
$M/G/1//N$, 289–291
$M/G/(a,b)/1$, 304–306
$M/G/c$, 322–324
  approximation for mean wait, 354
  diffusion approximation, 396–397
$M/G/c/c$, 322–324
$M/G/\infty$, 314–318, 333
  busy period, 333
  nonhomogeneous, 318
  steady-state distribution, 316
  transient distribution, 314–316
  with bulk arrival, 317
$M/M/\infty$, 83–84
$M/M/1$, 65–77
  average number in system, 67
  bulk arrivals, 174–178
  busy period, 119–124
  combination of service channels, 145
  distribution of number in system, 66–68
  output process, 72–74
  processor sharing, 245
  robustness, 74
  semi-Markov process analysis, 75–77
  service in random order, 71
  steady-state analysis, 66–68
  transient-state analysis, 111–119
  two-dimensional state model, 151–153
  waiting-time distribution, 68–71
$M/M/1/K$, 77–81
  busy period, 148
  equivalence with two-stage cyclic queue, 80–81
  steady-state solution, 77–78
$M/M/1/1$, 125–126
$M/M(a,b)/1$, 186–202
  busy period, 198–202
  idle period, 217

service batchsize distribution, 195–196
steady-state distribution, 187–190
transient-state distribution, 196–198
waiting-time distribution, 190–195
$M/M(a,b)/2$, 202–205
$M/M/c$, 84–95, 127–138
with balking and reneging, 154
busy period, 133–136
output process, 93–95, 133–136
steady-state distribution, 84–87
transient-state distribution, 127–131
waiting-time distribution, 89–93
$M/M(1,b)/1$, 201–202
$M/M(a,b)/2$, 202–205
steady-state solution, 203–204
transient-state distribution, 202–204
$M/M(1,b)/c$, 205–209
steady-state distribution, 208–209
transient-state distribution, 205–208
$M/M(a,b)/c$, 205–209
$M/M/c/c$, 95–100
blocking formula, 96–100
$M/M/\infty$, 84, 184
$M/M/c//m$, 101–110
application to electronics, 103
steady-state distribution, 101–102
waiting-time distribution, 104–106
$M^x/G/1$, 295–301
departure epoch system size, 295
waiting-time distribution, 295–301
$M^x/M/1$, 174–181
busy period distribution, 180–181
steady-state distribution, 175
transient-state distribution, 179–180
waiting-time distribution, 178–179
$M^x/M/\infty$, 181–185
steady-state distribution, 183–184
transient-state distribution, 181–183
Multiple Poisson bulk arrival, 212–213
Multiprogramming, level (degree) of, 236
Multiqueue, 138
Multiserver Poisson queue with ordered entry, 138–144

N

Network (of queues), 221–253
BCMP, 238–240
closed, 225
cyclic, 236–238, 245, 299
Gordon and Newell, 233–236
Jackson, 226–233
Normalization constant, 237
N-policy, 285–288, 370

O

Optimization models, 424
Ordered entry queue, 138–144
Output process, 72–73, 144
Overflow
discipline, 225
interoverflow time, 144
network, 144
system, 142–144

P

Pake's theorem, 444
Palm's integral equation, 144
PASTA, 59–62
Persistent state, 10
Perturbation rate, 74
Poisson
arrival process, 59–62
compound (cluster) process, 29–30
generalization of process, 29–31
nonhomogeneous process, 30–31
process, 25–34
properties of Poisson process, 28–29
role of Poisson process in probability models, 31
Pollaczek-Khinchin, 259–276, 296, 303, 327, 370
generalization of, 346–348
transform formula, 262
Processor sharing discipline, 50, 245

Q

Queue discipline, 49–50
blocking, 225
FCFS (FIFO), 49–50
LCFS (LIFO), 49–50, 257
loss, 219
overflow, 219
preemptive priority, 50
processor sharing, 50, 245
Queueing processes, 50–51
Queue (queueing system), 47–54
basic characteristics, 48–50
capacity of a, 49
cyclic, 237–238
general relationships, 54–58

Queue (queueing system) (*continued*)
non-Markovian, 255–338
notation, 51–52
transient and steady-state behavior, 52–54
with vacation(s), 398–423

R

Randomization, 32–34
Rate conservation law, 58
Rate equality principle, 65
Regenerative processes, 37–38
Renewal
alternating renewal process, 38
density, 35
function, 34
fundamental equation of theory, 36
process, 35–36
reward process, 44–45
superposition of processes, 59
theorem, 36
Residual lifetime, 36–37
Retrial queue (queueing system), 427–441
$M/G/1$, 432–436
$M/M/1$, 429–432
$M/M/2$, 437–438
$M/M/c$, 436–439
with finite orbit, 439–440
Rouché's theorem, 112–113, 119, 188
Rush hour, 383

S

Semi-Markov process, 39–41, 75–77
Service system
exhaustive, 398
gated, 408–412
limited, 407
nonexhaustive, 406–407
Stationary (invariant) distribution, 5–6
Steady-state behavior, 52–54
limitation of, 53–54
Stochastic decomposition, 399
Stochastic processes, 1–41
Sub-busy (pseudo) period, 277
Supplementary variable technique, 256, 267–274

T

Tandem queues (queues in series), 222–226
T-policy, 425
Traffic intensity, 50
Transient
behavior, 52–53, 110–112
state, 10
Transition density matrix, *see* Infinitesimal generator
Two-node system with feedback, 234–235

U

Uniformizable, 32
Utilization factor, 50

V

Vacation system (model), 398–423
with exhaustive service, 398–406
with gated service, 408–412
with limited service, 407–408
multiple, 399
with nonexhaustive service, 406–412
single, 399
variations, 414–415
Virtual waiting time, 125, 393

W

Weighing factor, 370
Wiener-Hopf
factorization, 368
integral equation, 342
technique, 309
Wiener process, 455
Work backlog, 391

Y

Yule Furry process, 25